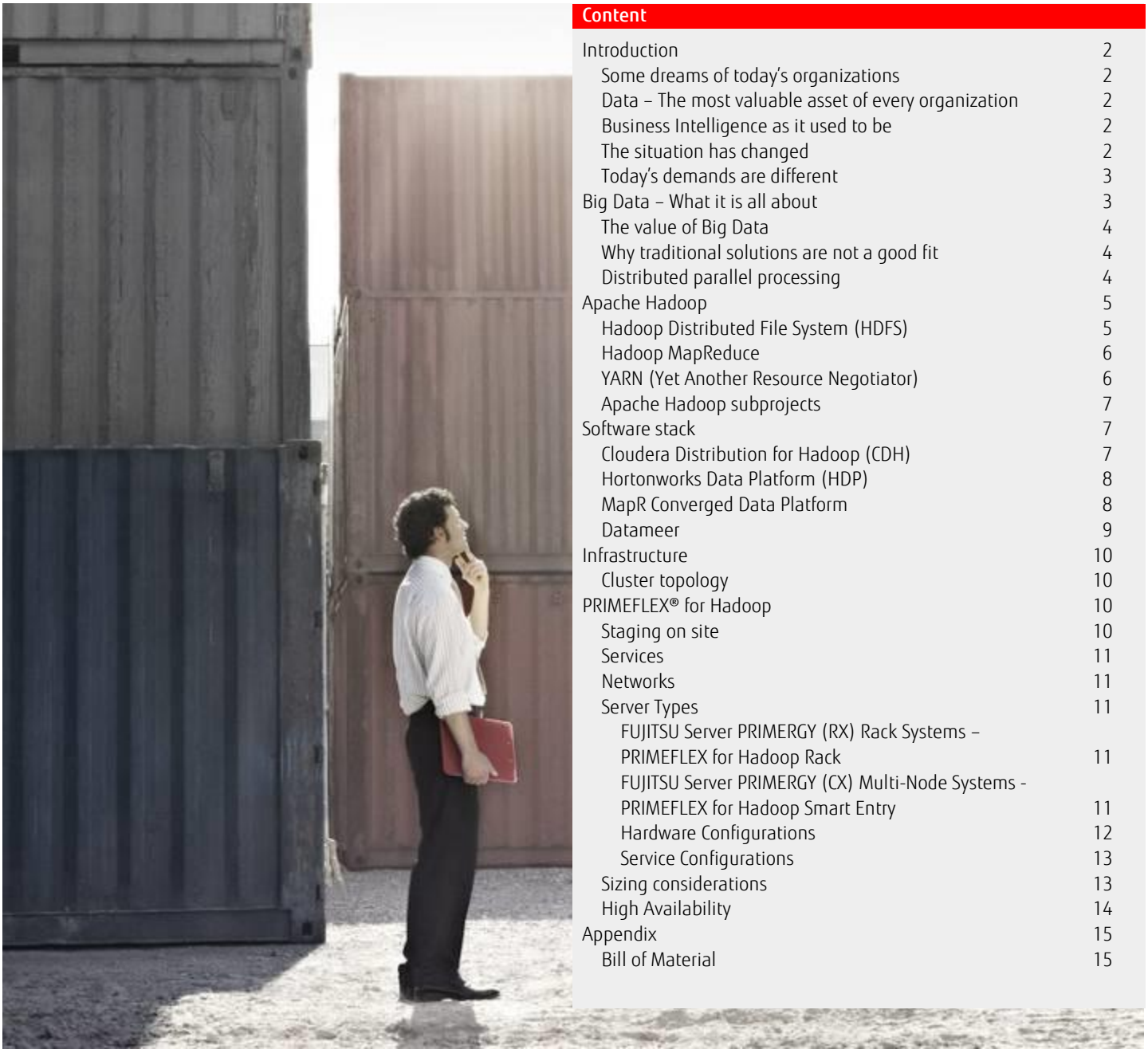
White paper
FUJITSU Integrated System PRIMEFLEX® for Hadoop

Reference Architecture based on either Cloudera Distribution for Hadoop, Hortonworks Data Platform or MapR Converged Data Platform and FUJITSU Server PRIMERGY, optionally including Datameer Analytics Software



Content

Introduction

Some dreams of today's organizations

Improving profitability and revenue is usually the top priority of an organization. This requires ever improving performance and productivity of their employees, efficiency, effectiveness and competitiveness of the overall business, while minimizing potential risks. The exciting question is how to achieve this faster, better and to a greater extent than your competitors.

- What if you could better predict what will happen in terms of trends, customer behavior, or business opportunities?
- What if you could always take the best decisions?
- What if you could accelerate your decisions making?
- What if you could have critical actions initiated automatically?
- What if you could fully understand the root cause of issues or costs?
- What if you could skip useless activities?
- What if you could quantify and minimize risks?

Contemplating such questions, many managers immediately imagine the opportunities for their business. However, are these only nice dreams, or is there a chance that these dreams can come true?

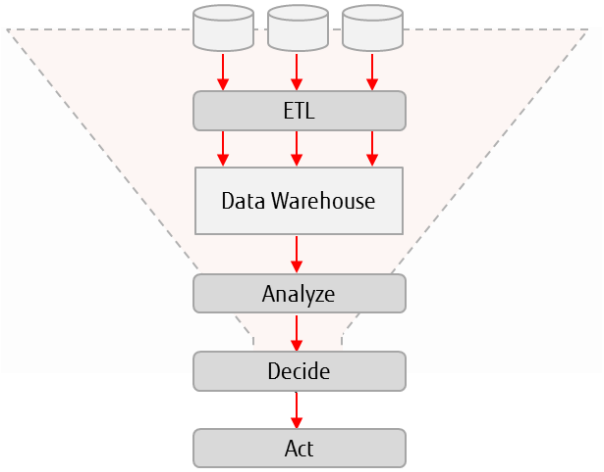
Data – The most valuable asset of every organization

Apart from human resources, data is the most valuable asset of every organization. Already decades ago, people were aware of this fact and tried to turn their data into value. It was obvious that utilizing data in an intelligent way for their business could support decisions based on real facts rather than intuition, thus helping improve business processes, minimize risk, reduce costs and increase business in general.

However, they also realized that data in its original form was usually of low value. Therefore data was collected from readily available data sources – mainly transactional databases – and then consolidated and transformed into a form convenient for analysis in order to discover relations, patterns and principles, to finally find the real value. Precisely this exactly was the idea of Business Intelligence (BI) in the early days.

Business Intelligence as it used to be

Typically in the context of Business Intelligence, the transformed data is loaded and stored in a dedicated database, the so-called Data Warehouse, which is separated from transactional systems in order to unload them from business analytics tasks, reports or visualization of query results in general. Data Warehouses are optimized for reporting.



For reasons of performance or authorization, multi-dimensional intervals or other special views are copied from the data warehouse; these so-called cubes or data marts can be used for in-depth analysis or role-aware reporting.

Traditional Business Intelligence considers mainly internal and historical views collected from a few data sources. Data is structured and typically stored in a relational database management system. Business analytics tasks are designed against a static data model, and happen periodically - every day, week or month in a batch process. As the average end user isn't trained to do his own sophisticated analysis, the number of direct users initiating queries or dealing with business analytics is strongly limited to a few specialists.

The situation has changed

Since the beginning of the Business Intelligence era, things have changed tremendously. There are versatile data sources which deserve to be taken into consideration. In addition to transactional databases, it is data from the web, be it blog contents or click streams which can help unveil valuable information, not to forget the content from social media which have evolved to the most commonly used communication platforms. There are multi-media files like video, photo and audio, from which important conclusions for the business can be drawn. There are huge text files including endless logs from IT systems, notes and e-mails which contain indicators that businesses are keen on. And not to forget the vast number of sensors built into smartphones, vehicles, buildings, robot systems, appliances, smart grids and whatever devices collecting data in a diversity which was unimaginable in the past. These sensors represent the basis for the ever evolving and frequently quoted Internet of things.

Having a closer look at specific industries, we should also mention medical scans in healthcare, RFID tags that track goods in motion, as well as geo-physical or spatial data (e.g. GPS-enabled location data) or data coming from observation satellites. This enumeration is far from complete.

It is a given, that every sort of data is constantly increasing in terms of volume, but especially sensors which generate event streams automatically and continuously, will have enormous influence. Therefore it is no surprise that we are all faced with an exponential data explosion.

Let us become more precise, what this exponential data explosion is all about. Analysts speak of 2.5 times 10^{18} bytes which are newly generated every day. 90% of all data originates from the last 2 years. There is a 65% growth of data per year, which equals a 100% growth every 18 months, or a 12-fold amount of data in 5 years compared to today. Consequently we are not just talking about Terabytes; we are talking about Petabytes, Exabytes, Zettabytes and even Yottabytes, and we do not recognize any real boundaries. As a result, many IT managers feel they are drowning in the vast floods of data.

It is not just the number of data sources and the amount of data that is increasing; it is also the types of data that proliferate. In the traditional Business Intelligence era, only structured data that reside in fixed fields of tables in relational database management systems was considered. Today, the majority of data – analysts speak of more than 80% - is unstructured. Unstructured data includes textual data, e.g. articles, e-mail and other documents, and non-textual data, e.g. audio, video and photos. In addition to structured and unstructured data, we should not ignore semi-structured data, which is not organized by fixed data fields, but has tags which separate consecutive data elements. Examples of semi-structured data are XML, HTML and PDF/A data, as well as RSS feeds.

Additionally to this, there is poly-structured data that includes a multitude of structures, which are even subject to change. Examples of poly-structured data are electronic records which represent XML documents with PDF/A elements, or different versions of the same document, which differ in the number of elements or even in the version of the underlying XML schema.

Today's demands are different

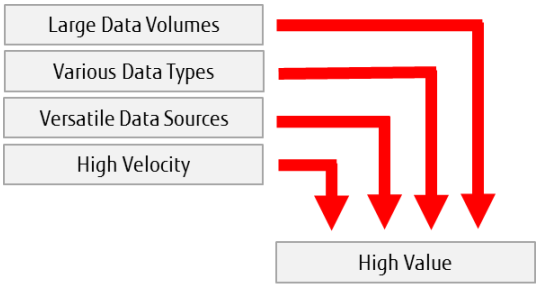
The exciting question is about the impact of all these considerations on Business Intelligence. People have recognized that within this vast amount of data from all the versatile data sources, no matter if structured, unstructured, semi-structured or poly-structured, which remained untapped in the past, much more value can be derived for organizations. But in contrast to traditional Business Intelligence, where reports were generated by batch processing within hours or days, ad-hoc queries with analytics results in real-time are now demanded in order to take immediate decisions proactively, or even initiate actions automatically. Moreover, the focus of analysis is on prediction about what will happen in the future rather than describing things which happened in the past. Due to the broad variety of opportunities all the data offers today, there are many more users who need direct access to business analytics, and this not only from the office but anywhere, from any device, be it a notebook, a smartphone or any other form factor.

Of course, a major requirement is that the solution to meet all these requirements has to be efficient and affordable.

With all this, we have set the stage for a new buzzword and one of the hottest topics and megatrends in today's IT market: Big Data.

Big Data – What it is all about

Big Data combines all the characteristics of data that we have just discussed. Big Data is defined by large data volumes in the range of many Terabytes and more – multiple petabytes is absolutely realistic, various data types (structured, unstructured, semi-structured and poly-structured data) from versatile data sources which are often physically distributed. Quite often, data is generated at high velocity and needs to be processed and analyzed in real-time. Sometimes data expires at the same high velocity as it is generated. From a content perspective, data can even be ambiguous, which makes its interpretation quite challenging.



However, Big Data is not just about the data itself; it is about affordable technologies to quickly store, discover and analyze massive data sets, sometimes even in real-time. High speed processing will enable you to submit queries repeatedly while gradually improving the queries and incrementally improving results. This enables many people, even those with little training, to be productive with analytics – something which would have been also absolutely unimaginable in the past.

In other words: Big Data stands for putting analytics and therefore knowledge at the fingertips of people that need it.

If you are wondering whether Big Data is a topic you should be interested in, there is a fairly simple answer. Just imagine that you, at the time being, utilize 5% of your available data for business analytics, which means in turn that 95% of your data is untapped. If you ignore Big Data and thus you are stuck at 5%, but your competitors from whom we assume that they use data with the same effectiveness are able to handle 15% of their data by using Big Data technologies, it will be pretty clear, which enterprise is more likely to win.

The value of Big Data

The value of Big Data for organizations is multifold. You may discover facts and insights about customers, suppliers and other business partners, about markets and operations, about the root causes of issues and costs, and the potential risks your organization might be faced with.

All these facts and insights would otherwise remain hidden. From newly discovered patterns and behaviors you will derive predictions about future trends and business opportunities which will definitely improve operational, tactical and strategic decision making in terms of speed, quality and relevance. Just avoiding some useless activities bears an enormous potential of cost savings. Big Data enables you to effectively use your data for business advantage and value creation. The option of initiating automatic actions will help accelerate this process.

Let us underline the value of Big Data using some examples. The impact of new insights can be an innovation of your business, your products and / or services. Customers who are likely to churn can be retained, and churned customers will be won back, by reliably evaluating customer sentiments, e.g. from the state of delivery processes and their calls to the helpdesk. New customers will be attracted by identifying current demand, e.g. through social media analysis. At the same time, a better-targeted marketing campaign will yield a higher ROMI (Return of Marketing Invest). Other examples are closely related to improved business processes. Examples are reducing search and processing time, achieving better planning and forecasting by a detailed analysis of historical data, improving the allocation of physical and human resources or increased performance and productivity by automated decision making in real-time. Finally, improvement of efficiency and effectiveness will increase profitability and drive growth. Business will benefit tremendously by being able to quantify and minimize risk. Capitalizing on information will help improve your competitiveness.

Why traditional solutions are not a good fit

As mentioned before, Data Warehouses are the data stores of traditional Business Intelligence solutions. Typically, they are based on relational databases. Relational databases require some sort of structure, which means that unstructured and semi-structured data would need some pre-processing. The tables resulting from pre-processing are often huge but sparse. This in turn means a high amount of metadata, high storage capacities and slow access to data. A row-wise data organization is well-suited for online transaction processing (OLTP), but when it comes to analytical tasks, much irrelevant data will have to be read from the large number of rows, because only information in certain columns is relevant.

Can server scale up improve the situation? No matter how powerful your server might be, there will always be a hard limit for each resource type. With increasing data size, these limits will become a challenge, sooner or later. For sure, in the future the limits will move upwards, but the totality of all data involved in your analysis might grow much faster. Moreover, with high-end servers and scale-up, cost for CPU, main memory and networking will always be relatively high.

If scale up is not recommended, the question remains about relational databases and server scale out. As the database is shared among several servers and accessed by them, the storage connections can prove as a bottleneck. Likewise, the effort to coordinate the access to shared data will increase with the number of database servers. This will – according to Amdahl's law - lead to decreasing efficiency of the individual server, and will strongly limit parallelization. Amdahl's law says that with every parallelization approach the inverse of the sequential portion which cannot be executed in parallel will limit the theoretically achievable performance increase. Hence, in order to achieve a 1000-fold performance increase by using 1000 server nodes, a serial portion of less than 0.1% has to be striven for.

Consequently, everything you try to improve the situation, be it scale up or scale out in conjunction with your relational database, would be time-consuming and expensive and far away from fulfilling real-time demands. Results would be returned too late and insights could become obsolete when being presented to the user. All told, due to the high volume of data, relational databases will outgrow economic feasibility and the required performance is unachievable.

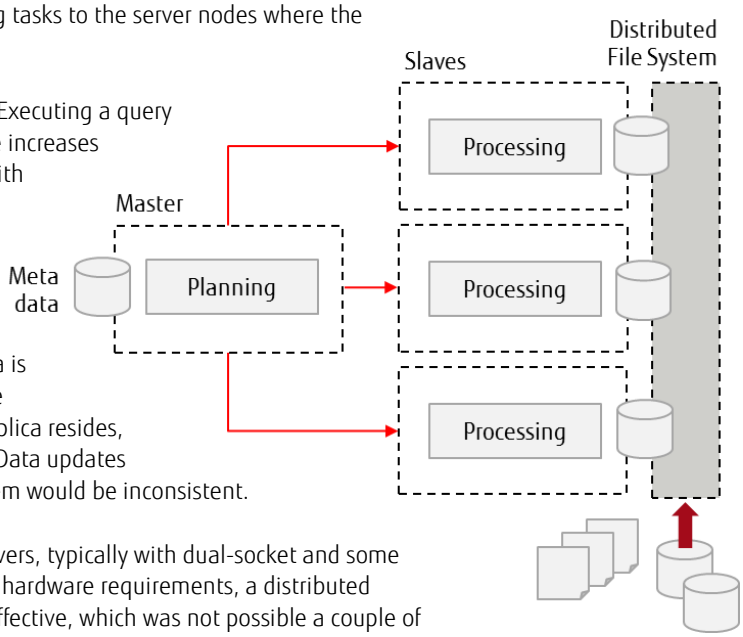
For sure, you could think about building distinct databases and Data Warehouses and split up your analytics tasks. However doing so, you would create disconnected data silos and conduct distinct disconnected analytics which would not give you the comprehensive insight that you expect.

As you would experience limitations everywhere when using traditional solutions, new approaches are required, which keep processing time constant, while data volumes increase.

Distributed parallel processing

The key to success is parallelization based on a "shared nothing" architecture and non-blocking networks ensuring a smooth communication among servers. You distribute I/O to many server nodes by moving subsets of data to the local storage of the servers.

Likewise, data processing is distributed by moving computing tasks to the server nodes where the data resides.



Distributed parallel processing provides several advantages. Executing a query or any other data operation by many nodes at the same time increases performance and delivers fast results. You may start small, with just a few servers, and you add more servers, as they are needed. Basically, your infrastructure will linearly scale out without any limits. Even if 1000s of servers are not enough, the game of adding new ones remains the same.
In order to make the overall configuration fault-tolerant, data is automatically replicated to several nodes. If a server fails, the respective task can be continued on a server where a data replica resides, fully transparently for software development and operation. Data updates need to be considered for all data copies, otherwise the system would be inconsistent.

As such a server farm can be built from industry standard servers, typically with dual-socket and some terabytes of local storage, but without any special additional hardware requirements, a distributed parallel processing solution will prove to be extremely cost-effective, which was not possible a couple of years ago. On the software layer the server farm is coordinated as a distributed storage system which coincidentally runs as a parallel processing cluster. This is seen as one of the biggest enablers for Big Data analytics.

Apache Hadoop

The de-facto standard for Big Data and distributed parallel processing is Hadoop, an open source framework primarily for batch operation, written in Java. Hadoop is a project and registered trademark of the Apache Software Foundation. It is designed to scale up to 1000s of nodes, to accept server crashes as "normal" in large farms, and to make data storage and analytics robust against failures.

The core components of Hadoop are the Hadoop Distributed File System (HDFS) and the Hadoop MapReduce framework.

Hadoop Distributed File System (HDFS)

HDFS is a distributed file system which is in particular designed and optimized for processing large data volumes and for highest availability. It spreads across the local storage of a cluster consisting of many server nodes.

HDFS includes a master server (NameNode) that partitions the original data and assign the data to the server nodes, based on defined rules. Each slave node of the cluster stores just a small fragment of the complete data set. The DataNode which is installed on each slave node is in charge of the local data management.

Each data block is replicated on more than one server for the purpose of high availability. By default, every data block exists three times. Besides the primary data block, one copy typically exists on a server in the same rack, while an additional copy will be on a server in another rack. To increase availability even more, data can be distributed to different locations. Of course, this will not protect against human errors when copying or deleting data; for this purpose, additional backup processes need to be applied.

The HDFS NameNode manages the metadata, thus being always aware of which data blocks belong to which files, where the data blocks are located, and where which storage capacities are occupied. By means of periodically transmitted signals, the NameNode will always know which DataNodes are still working. If the signal is missing, the NameNode will recognize the failure of a DataNode, will remove the failed DataNode from the Hadoop cluster, and will always try to distribute data load evenly across the available DataNodes. Furthermore, the NameNode ensures that the defined number of data copies is always available.

The HDFS API defines the contract between HDFS implementation and HDFS clients. A Hadoop distribution may replace the above described standard implementation with a custom implementation of the HDFS API.

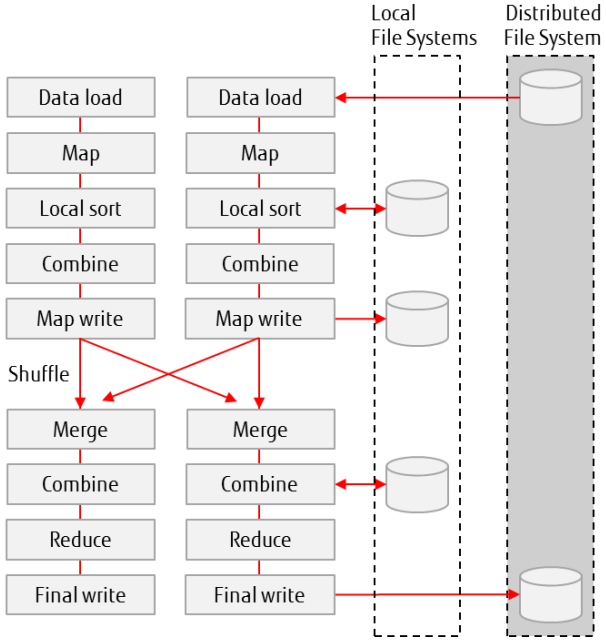
Hadoop MapReduce

Similar to HDFS, the MapReduce framework works according to the master-slave principle. The master divides a given problem (job) into multiple tasks (map tasks), and distributes these tasks across the network to a number of slave nodes for parallel processing.

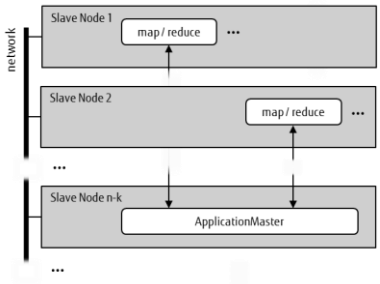
Typically, the map tasks run on the same cluster nodes, where the processed data resides. If that server node is already heavily loaded, another node will be selected which is close to the data, i.e. preferably a node in the same rack.

Intermediate results will be exchanged among the nodes (shuffling), and thereafter merged by the reduce tasks to a final result. As the processing tasks are moved to the data and not vice versa, in the map phase basically all I/O activities can be parallelized while network load is almost completely avoided. To avoid any bottleneck regarding scalability in the shuffle phase, a non-blocking switched network without or with low overprovisioning between the server nodes is required.

Optionally, intermediate results of the map phase and the shuffle phase may be aggregated (combine), in other to keep data volumes transferred from map tasks to reduced tasks as low as possible.



While the input data for MapReduce as well as final results reside in HDFS, the intermediate results are deposited in the local file systems of the DataNodes.



The master role in the MapReduce framework is assumed by the ApplicationMaster. It starts the map and reduce tasks and monitors their progress on all nodes and takes care that disrupted or aborted tasks will be executed anew.

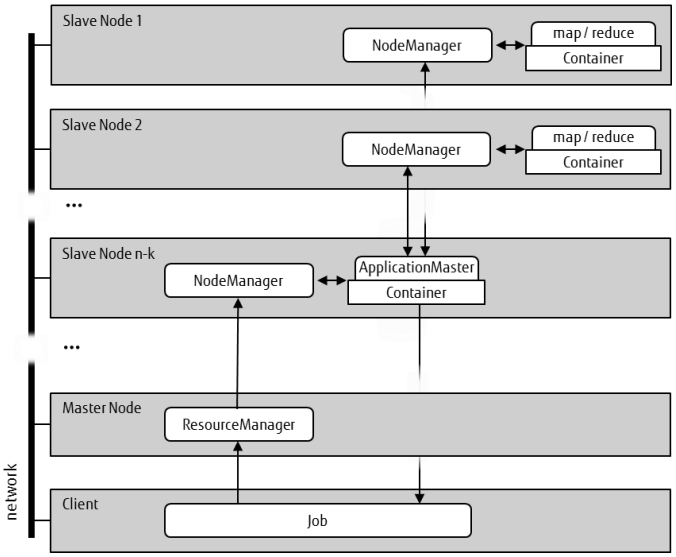
If a task does not notify any progress for a longer period, or if a slave node fails completely, all tasks not terminated yet will be restarted on another server, usually on a server with a respective copy of the data. If a task runs extremely slowly in comparison with the other tasks, the ApplicationMaster will also start a second instance of the task on another server in order to execute the overall job in good time (speculative execution).

It is true that MapReduce can be directly applied for executing business analytics. But a frequent use case is to transform data into an optimized shape for analytics.

YARN (Yet Another Resource Negotiator)

YARN is a further development of the initial MapReduce framework. Cluster resource management and application control which in MapReduce v1 are both covered by a single component, are split up into separate components. It is just the lean ResourceManager that still exists as a central instance, while the entire job control is delegated to the ApplicationMaster which can run on any slave node. For every job an ApplicationMaster is dynamically created; it is exclusively available for this job. MapReduce v2 will then only look after the parallel data processing. An ApplicationMaster sends periodic heartbeats to the ResourceManager, and in the event of ApplicationMaster failure, the ResourceManager will detect the failure and start a new instance of the master running in a new container. On each slave node, a NodeManager is responsible for local resource management of containers.

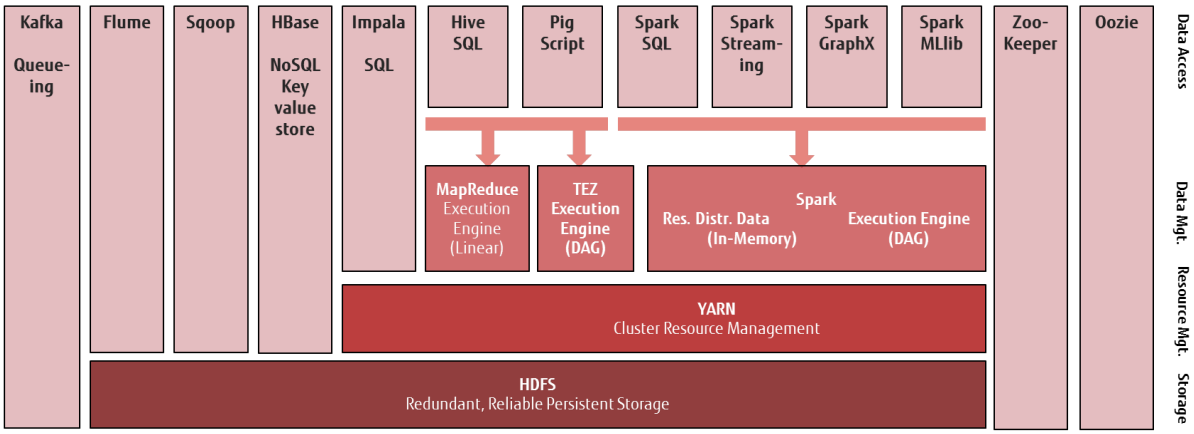
YARN is a general framework which can run MapReduce jobs and other kinds of application.



The distributed application control increases the level of parallelization, removes bottlenecks and enables clusters with 10,000s of nodes.

Apache Hadoop subprojects

Beside the core components **HDFS**, **YARN** and **MapReduce**, there are various open source subprojects under the Apache license, which make Hadoop a comprehensive platform for analytic applications. A selection of the most prominent function blocks and their dependencies is shown in the adjacent picture.



Flume is a reliable distributed system for collecting, aggregating and moving large amounts of log data. It is suited in particular for importing log data streams into HDFS.

By means of the NoSQL database **HBase** large tables can be stored and accessed efficiently.

Hive and the declarative query language HiveQL form a data warehouse used for ad-hoc queries and simplified report generation. The compilation into MapReduce jobs happens automatically. Beside others Hive can process data form plain text files or HBase format.

Impala provides high-performance, low-latency SQL queries on data stored in popular Apache Hadoop file formats.
It integrates with the Apache Hive metastore database, to share databases and tables between both components. Impala can access data directly from the HDFS file system and it also provides a SQL front-end to access data in the HBase database system.

Kafka is a high-throughput distributed publish-subscribe messaging system. It supports partitioning messages over Kafka servers and distributing consumption over a cluster of consumer machines while maintaining per-partition ordering semantics. Kafka supports parallel data load into Hadoop.

Workflows can be described and automated using **Oozie** by considering the dependencies between individual jobs.

Pig with the script language „Pig Latin" enables the creation of scripts for queries and reports, which are compiled into MapReduce jobs.

Using **Spark**, you have the option to load data from HDFS or HBase into the memory of the cluster nodes for faster processing. A core feature of Spark is the effective management of in-memory data containers. This is of particular advantage, if multiple queries are applied to the same data sets or computation results serve as input for a next step, as for instance with machine learning algorithms. Currently there are four APIs leveraging Spark technology: **Spark SQL**, **Spark Streaming**, **Spark MLlib** and **Spark GraphX**. They deliver a datawarehouse functionality like Hive, an event processing service, a machine learning library like Mahout and a graph computation library.

Sqoop is used for loading large data volumes from relational databases into HDFS and vice versa.

ZooKeeper is a library of modules for implementing coordination and synchronization services in a Hadoop cluster.

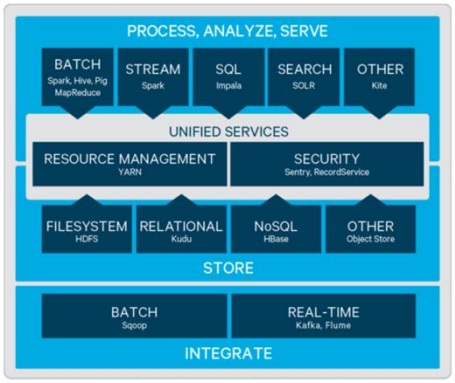
Software stack

PRIMEFLEX for Hadoop offers a choice of three major Hadoop distributions:
Cloudera Distribution for Hadoop, Hortonworks Data Platform or MapR Converged Data Platform.

Cloudera Distribution for Hadoop (CDH)

CDH is the world's most complete, tested, and popular distribution of Apache Hadoop and related projects. CDH is 100% Apache-licensed open source and is the only Hadoop solution to offer unified batch processing, interactive SQL, and interactive search, and role-based access controls.

CDH includes the core elements of Apache Hadoop plus several additional key open source projects that, when coupled with customer support, management, and governance through a Cloudera Enterprise subscription, can deliver an enterprise data hub.



Source: https://www.cloudera.com/documentation/enterprise/latest/topics/cdh_intro.html

Cloudera provides three closed-source components with enterprise-grade functionality: Cloudera Manager, Cloudera Data Science Workbench and Cloudera Navigator.

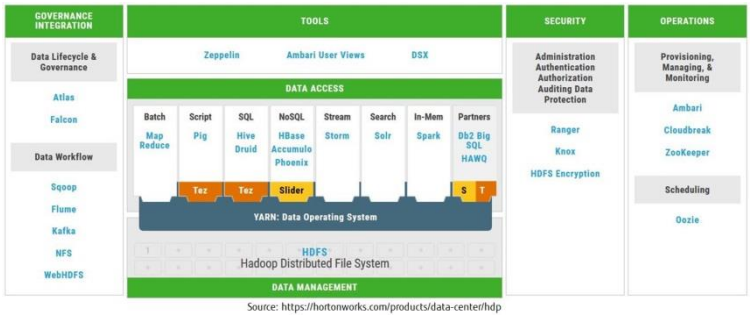
Cloudera Manager is the industry's first and most sophisticated management application for Apache Hadoop and the enterprise data hub. Cloudera Manager sets the standard for enterprise deployment by delivering granular visibility into and control over every part of the data hub – empowering operators to improve performance, enhance quality of service, increase compliance and reduce administrative costs. Cloudera Manager is designed to make administration of your enterprise data hub simple and straightforward, at any scale. With Cloudera Manager, you can easily deploy and centrally operate the complete Big Data stack. The application automates the installation process, reducing deployment time from weeks to minutes; gives you a cluster-wide, real-time view of nodes and services running; provides a single, central console to enact configuration changes across your cluster; and incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization.

Cloudera Data Science Workbench (CDSW) enables fast, easy, and secure self-service data science for the enterprise. It allows data scientists to bring their existing skills and favorite tools, such as R, Python, and Scala, to securely run computations directly on data in Hadoop clusters. CDSW provides an environment for Deep Learning Frameworks. It is a collaborative, scalable, and extensible platform for data exploration, analysis, modeling, and visualization. CDSW lets data scientists manage their own analytics pipelines and share insights with their team, thus accelerating machine learning projects from exploration to production. Optional GPUs can boost performance for applications in CDSW.

Cloudera Navigator is the only complete data governance solution for Hadoop, offering critical capabilities such as data discovery, continuous optimization, audit, lineage, metadata management and policy enforcement.The Cloudera product is available with different feature sets. For a comparison please refer to https://www.cloudera.com/content/dam/www/marketing/resources/datasheets/cloudera-enterprise-datasheet.pdf.landing.html

Hortonworks Data Platform (HDP)

HDP is the industry's only true secure, enterprise-ready open source Apache™ Hadoop® distribution based on a centralized architecture with YARN in the middle as the leading component.



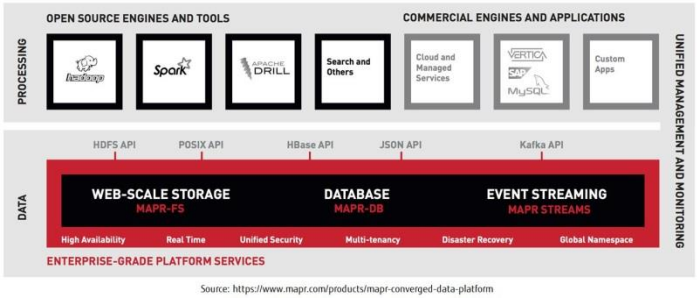
Source: https://hortonworks.com/products/data-center/hdp

HDP addresses the complete needs of data-at-rest, powers real-time customer applications and delivers robust analytics that accelerate decision making and innovation.

The Hortonworks product contains multiple components. For detailled information please refer to http://hortonworks.com/products/data-center/hdp

MapR Converged Data Platform

The MapR Converged Data Platform integrates Hadoop, Spark, and Apache Drill with real-time database capabilities, global event streaming, and scalable enterprise storage to power a new generation of big data applications.



Source: https://www.mapr.com/products/mapr-converged-data-platform

The MapR Platform delivers enterprise grade security, reliability, and real-time performance while dramatically lowering both hardware and operational costs of your most important applications and data.

MapR-FS is a POSIX file system that provides distributed, reliable, high performance, scalable, and full read/write data storage. MapR-FS supports the HDFS API and fast NFS access. Metadata, which is kept by the NameNode in traditional HDFS, is distributed over all cluster nodes with MapR-FS, thus eliminating the central NameNode. MapR-FS stores data in abstract entities called containers that reside on storage pools. The Container Location Database (CLDB) is a light-weight MapR-FS component which maintains information about the location of every container in the cluster, defines the container precedence in the replication chain, and organizes container content updates across the replication chain.

The MapR product is available with different feature sets. For a comparison please refer to https://www.mapr.com/products/mapr-distribution-editions

Datameer

Datameer, an optional component of PRIMEFLEX for Hadoop, is the only end-to-end big data analytics application purpose-built for Hadoop that enables the fastest time from raw data to new insights.

- Integrate All Your Data

You have a lot of data, in lots of different formats, in lots of different locations. Some of it is in the cloud, some of it is in various legacy databases, some of it is in spreadsheets on your desktop. You've got big data. Hadoop can help, but you don't speak Hadoop, and you don't have the time or resources to find someone who does. That's where Datameer comes in. Datameer makes integrating all your data into Hadoop as easy as following a wizard.

With built-in connectors to all common structured and unstructured data sources, big data integration is an easy, three step process of where, what and when. Once you tell Datameer what data you want brought in to Hadoop, all you need to do is decide how you want it brought in, whether it's a one-time import or streamed in as new data is added or on a schedule determined by you. Say goodbye to ETL and pre-defined schemas. This is the new data integration reality for Hadoop big data analytics.

- Analyze All Your Data – Smarter and Faster

With Datameer, big data analytics is as simple as using a spreadsheet. Our data integration wizards make it easy to integrate all of your data into Hadoop. And once it's in Hadoop, there's no limit to what you can do with analytics. We don't use pre-defined schemas, so you can ask any question about any data at any time. Building an analysis is as simple as selecting which data to work with in a spreadsheet, choosing from over 250 pre-built analytic functions available in Datameer, and iterating at the speed of thought, using Datameer's patented Smart Sampling™ technology. This flexibility and ease of use makes Datameer the ideal application for big data discovery in Hadoop.

When you're ready to run your workbook, Datameer's patent pending Smart Execution™ technology intelligently and dynamically selects and combines the computation frameworks MapReduce, Tez and Spark to achieve maximum performance for your analytics task. Smart Execution™ improves performance and cluster utilization by selecting the most appropriate and efficient compute framework for each step of your analysis, and using in-memory technology to process smaller data sets.

At any time, you can click to view the "flip side" of your spreadsheet for an instant, visual profile of your data. This allows you to instantly detect data quality issues, understand the effect of every transformation you make, adjust your analysis as you go, and ensure there are no surprises in your final visualization.

- Visualize All Your Data

Communicating the results of your Hadoop big data analysis is easy in Datameer's WYSIWYG Business Infographic Designer. Instead of being constrained to box after box of charts and graphs, Datameer provides you with a blank HTML5 canvas to design beautiful infographics or reports that will automatically update every time your data updates. You can import any image, embed a video, write free-form text, and customize to no end. Even better, thanks to HTML5, your visualizations are consumable on any device. Gone are the days when you had to take a screenshot of your dashboard and annotate in PowerPoint.

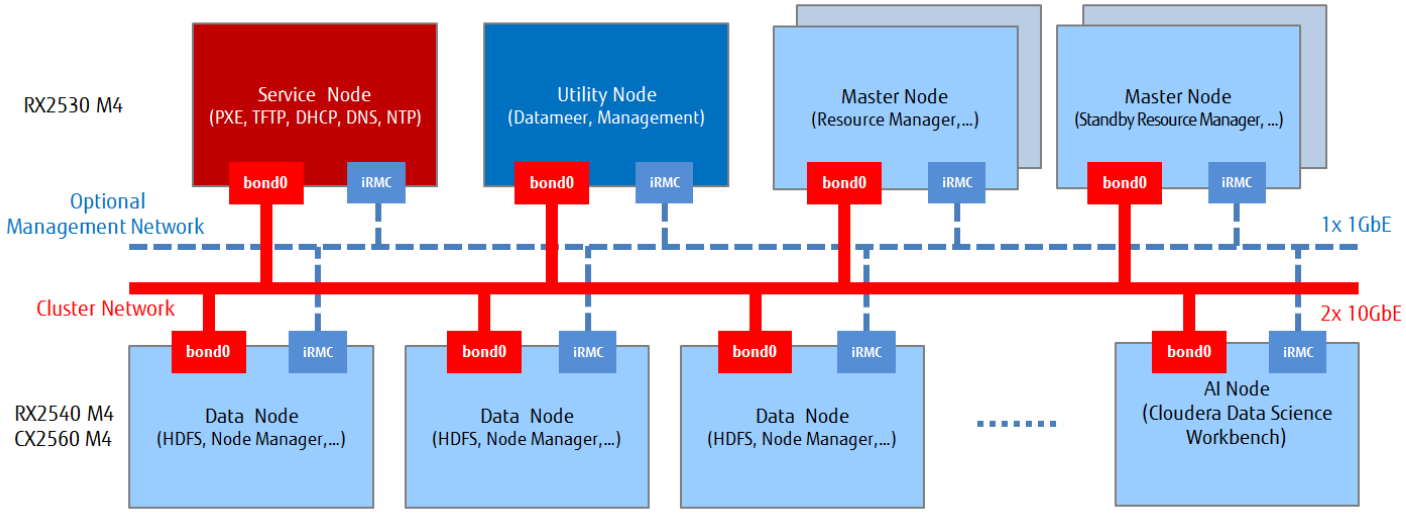
For more information about Datameer please refer to http://www.datameer.com/

Infrastructure

The considered Big Data software stack consisting of a Hadoop execution and data management layer and the optional analytical software from Datameer runs on a distributed scale out server infrastructure. The Hadoop layer is provided by one of the following Hadoop distributions: Cloudera Distribution for Hadoop, Hortonworks Data Platform or MapR Converged Data Platform. The infrastructure is built by servers in different roles and a powerful network. The storage capacity is provided by direct attached disks inside the servers linked together to a distributed file system. There is no external storage system as part of the infrastructure.

Cluster topology

All nodes are connected to the Cluster Network with two bonded 10GbE interfaces (bond0). The cluster network is connected to the datacenter network through Spine-Leave switch architecture. The remote management boards can be connected by a separate 1GbE network or use the shared 10GbE interface.



The **Service Node** provides all installation service for the initial setup of all other servers. It can also be used as DNS server and NTP server if these services are not available in the customer network.

Datameer and management components of the Hadoop distribution share one node. This node is the central access point for the cluster providing the web interfaces for analytics and cluster management.

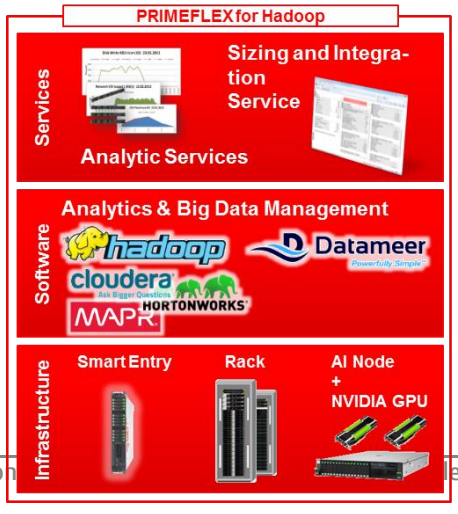
Master Nodes can use each other as a standby node, e.g. for the YARN **Resource Manager**. In large configurations these standby nodes can also run on additional nodes in different racks.

The **Data Nodes** do the real work in the cluster. They contain the disks used for HDFS and run the YARN Node Manager for performing analytics tasks on the data.

AI Nodes host the Cloudera Data Science Workbench for data exploration, analysis, modeling and visualization.

PRIMEFLEX® for Hadoop

PRIMEFLEX for Hadoop is a powerful and scalable platform as an integrated system that provides business users with a more cost-effective way of creating actionable analytics from big data. PRIMEFLEX for Hadoop analyses large volumes of data to extract and make accessible meaningful business-relevant information, combining the convenience of pre-configured and pre-tested hardware and the economic advantages of open source software plus system support and all-round lifecycle management. For worry-free operation, strategic big data consulting, analytics consulting, consulting for Hadoop, and integration and maintenance services are supplementing the offering which helps to create knowledge from information.



Staging on site

PRIMEFLEX for Hadoop comes with preinstalled software including RedHat Enterprise OS, Datameer and one of the supported Hadoop distributions. The Entry variant is completely installed and configured in the factory and needs only to be connected to the customer network.

The rack variants are fully staged with the software in the factory if standard racks are used. If customer specific racks are contained the staging is done from the ready-to-run service node

at the customer site after building the racks in the data center. The cluster setup with the preinstalled Hadoop management components and the configuration of Datameer is always done at the customer by Fujitsu service.

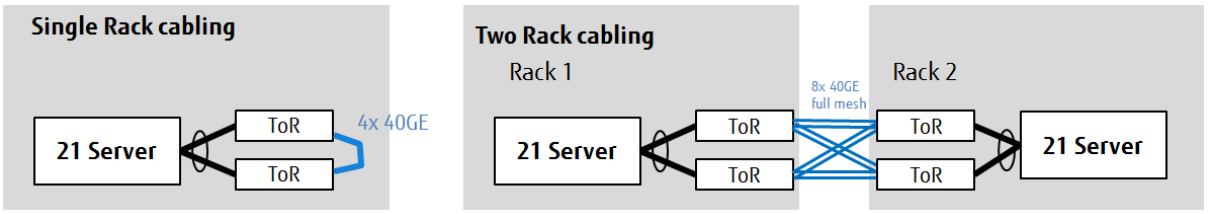
Services

Fujitsu offers end to end integration and consulting services for PRIMEFLEX for Hadoop, from strategic consulting to use case development and implementation services for specific customer needs. Please consult your local Fujitsu representative to see what services are available in your region.

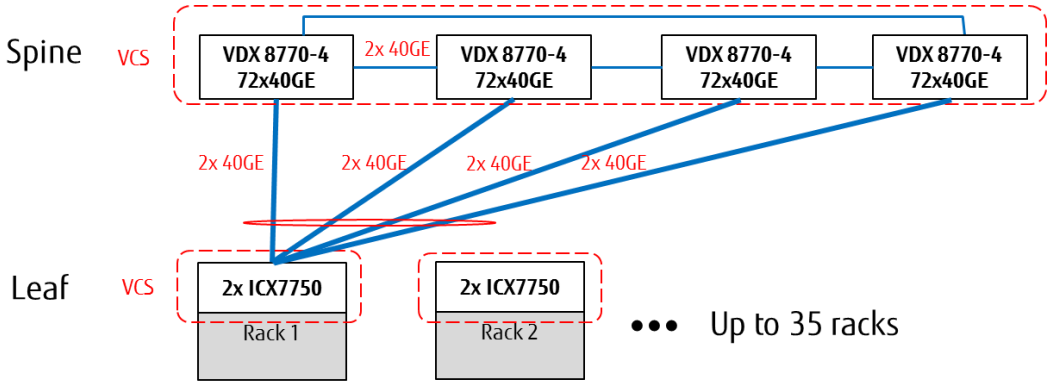
Networks

Sufficient network capacity is crucial for a performant Hadoop cluster. So the cluster interconnect in PRIMEFLEX for Hadoop is always built on 10GbE network components. The selection of the components depends on the existing customer networking infrastructure and can be provided from latest Brocade or Cisco switching technologies.

Fujitsu offers two configurations for different requirements: PRIMEFLEX for Hadoop Rack is intended for high storage capacity and PRIMEFLEX for Hadoop Smart Entry targets TODO. The following paragraphs describe the Rack configuration.



The rack systems come with up to 2 top of rack switches. Each node is connected to them with 2 x 10GbE. Additionally up to 2 1GbE switches are in the racks for management and deployment. There can be up to 21 servers in one rack. In the following example the 2 ToR switches can be Brocade VDX 6740 or Cisco Nexus 56128P.
For clusters containing more than two racks spine-leaf network architecture is recommended. Here is an example for PRIMEFLEX for Hadoop Rack with up to 21 servers per rack. Each of the 2 VDX6740 ToR switches in each rack is connected to 2 VDX8770 spine switches by one 40GbE cable. So the total connection bandwidth from one rack to the spine layer is 8x40=320Gb. The 21 servers in one rack are connected with 2x10GbE to the ToR switches which gives a total of 21x10x2=420Gb. This results in a 1.3:1 over commitment of the leaf/spine bandwidth which is a very good value and assures a 10Gb connection speed all over the whole cluster.



Server Types

FUJITSU Server PRIMERGY (RX) Rack Systems – PRIMEFLEX for Hadoop Rack

PRIMERGY Rack Servers are the perfect platform to build up dynamic infrastructures to support current and future business requirements by delivering top performance, excellent scalability and leading efficiency in a modular, easy-to-upgrade housing. Combined with high efficient power supplies and a simplified power management this contributes to favorable lifecycle costs. The PRIMERGY RX2540 server nodes with up to 408 cores and over 1PB of disk storage per rack address the deployment of platforms for storage-intensive big data analytics.

FUJITSU Server PRIMERGY (CX) Multi-Node Systems - PRIMEFLEX for Hadoop Smart Entry

High compactness, a shared power and cooling infrastructure and ease of use aiming at decreased capital and operational expenses are central design characteristics of the PRIMERGY CX400 architecture. As part of the PRIMERGY CX multi-node server line, it provides a

new condensed granularity for scale-out server deployment by packing up to 4 half-wide dual socket server nodes into an only 2 U chassis in standard 19" rack form factor. PRIMEFLEX for Hadoop Smart Entry is a ready-to-run appliance with 64 cores / 48 TB.

Hardware Configurations

These are only the standard configurations which can be adapted to special customer needs.

Datameer and Hadoop Management Node in Rack Configuration	
PRIMERGY Server	**RX2530 M4**
CPU	2 x Intel Xeon Silver 4110 8C 2.10 GHz
RAM	4 x 16GB (1x 16GB) 1Rx4 DDR4-2666 R ECC
Disk	6 x SSD SATA 6G 240GB Mixed-Use 2.5" H-P EP
Network	PLAN EM 2x 10Gbit SFP+ OCP interface

Master Nodes in Rack Configuration	
PRIMERGY Server	**RX2530 M4**
CPU	2 x Intel Xeon Gold 6134 8C 3.20 GHz
RAM	8 x 32GB (1x 32GB) 2Rx4 DDR4-2666 R ECC
Disk	6 x SSD SATA 6G 240GB Mixed-Use 2.5" H-P EP
Network	PLAN EM 2x 10Gbit SFP+ OCP interface

Each Master Node is equipped with 6 disks which are configured for different purposes:

- 2 x RAID 1 OS Root disk
- 2 x RAID 1 HDFS Namenode metadata
- 1 x RAID 0 Zookeeper
- 1 x RAID 0 Quorum Journal Node

Data Nodes in Rack Configuration / all nodes in Smart Entry Configuration		
PRIMERGY Server	**RX2540 M4**	**CX2560 M4**
CPU	2 x Intel Xeon Silver 4110 8C 2.10 GHz	2 x Intel Xeon Silver 4110 8C 2.10 GHz
RAM	8 x 16GB (1x 16GB) 1Rx4 DDR4-2666 R ECC (max. TODOGB)	8 x 32GB (1x 132GB) 2Rx4 DDR4-2666 R ECC
Disk	12 x HD SAS 12G 4TB 7.2K 512e HOT PL 3.5" BC or 24 x HD SAS 12G 2TB 7.2K 512e HOT PL 2.5" BC 4 x SSD SATA 6G 240GB Mixed-Use 2.5" H-P EP	6 x HD SATA 6G 2TB 7.2K 512n HOT PL 2.5' BC
Network	PLAN EM 2x 10Gbit SFP+ OCP interface	10 GBE Cross Connections

The disks of the RX2540 are configured as follows:

- 4 x RAID 5 (SSD SATA HDs) OS disks
- 12 or 24 x RAID 0 distributed file system disks

The disks of the CX2560 are configured as follows:

- 1 x RAID 0 OS disks
- 5 x RAID 0 distributed file system disks

If the Cloudera distribution is selected, then AI Nodes can additionally be configured for the Rack configuration:

AI Node	
PRIMERGY Server	**RX2540 M4**
CPU	2 x Intel Xeon Gold 6130 16C 2.10 GHz
RAM	6 x 32GB (1x 32GB) 2Rx4 DDR4-2666 R ECC
Disk	8 x SSD SAS 12G 400GB Mixed-Use 2.5" H-P EP
Network	PLAN EM 2x 10Gbit SFP+ OCP interface
GPU	2 x PY NVIDIA Tesla P100 for PCIe 12GB

If the Cloudera or Hortonworks distribution is selected, then the hardware configuration of the Master and Data Nodes will be as described above.
A standard hardware configuration for the MapR distribution does not require dedicated Master Nodes; all nodes are configured in the same way as RX2540 Data Nodes. Dedicated RX2530 Master Nodes as described above are only recommended for very large configurations with hundreds of nodes.

Service Configurations

Here is an example how the services may be distributed between the cluster nodes for CDH:

Node / Service	Utility Node	Master Node 1	Master Node 2	Data Nodes	AI Nodes
Management	Cloudera Manager, Cloudera Agent, Management Services	Cloudera Agent	Cloudera Agent	Cloudera Agent	Cloudera Agent
ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper		
HDFS	QJN	NN,QJN	NN (Standby),QJN	Data Node	Gateway
YARN		RM (Standby), History Server	RM	Node Manager	Gateway
Datameer	Datameer				
Workbench					Cloudera Data Science Workbench
Hive	MetaStore		WebHCat, HiveServer2		
Oozie	Oozie				
Navigator			Navigator, KMS		
HUE	Hue				
SOLR				search	
Spark					Gateway
impala	Catalog, Statestore			impalad	
HBASE		HMaster	HMaster	Region Servers	

Here is an example how the main services may be distributed between the cluster nodes for HDP:

Node / Service	Utility Node	Master Node 1	Master Node 2	Data Nodes
Management	Ambari Server	Ambari Agent	Ambari Agent	Ambari Agent
ZooKeeper	ZooKeeper (only 1,3,..)	ZooKeeper	ZooKeeper	
HDFS		NN	NN (Standby)	Data Node
YARN		RM (Standby), HistoryServer	RM,	Node Manager
Datameer	Datameer			

Here is an example how the main services may be distributed between the cluster nodes for MapR:

Node / Service	Utility Node	Data Node 1	Data Node 2	Data Node N
Management		Warden	Warden	Warden
ZooKeeper		ZooKeeper	ZooKeeper	ZooKeeper (only 1,3,..)
Storage		CLDB, File Server	CLDB, File Server	CLDB (0, 1), File Server
YARN		RM (Standby)	RM	Node Manager, HistoryServer (1)
Datameer	Datameer			

Sizing considerations

Suitable sizing of all components in a Hadoop cluster is essential for smooth operation. The scope of sizing not only covers individual master and slave nodes with their resources CPU, RAM, disks and network interfaces but also the cluster size and its interconnects.

Major influencing factors for sizing a Hadoop cluster are data volume, compression ratios, workload performance requirements and availability requirements. Workloads may be executed with different kinds of computation frameworks, e.g. MapReduce, Spark or Tez, and each framework has different requirements depending on the way it works.

- Clusters are typically sized to accommodate a growing volume of data for a given period of time. Data includes clearly visible user data residing in HDFS, be it permanently or temporarily, as input to or output from jobs of many computation frameworks, and less visible intermediate data from MapReduce jobs which is temporarily stored in local file systems on each slave node.
- RAM needs to be sized according to workload requirements. This is of special importance for workloads run with Spark and Tez which use in-memory technology in order to avoid the overhead of storing data on disk between workload steps.
- Compression may have a significant impact on the sizing requirements depending on the compression factor of the selected compression mechanism and type of data. Compression can be applied independently to data in HDFS and to intermediate MapReduce data.
- Workloads in a Hadoop cluster can be broken down into consecutive and concurrent jobs where each job has its own characteristics, i.e. CPU or I/O intensive. For the largest jobs, analyze performance requirements for all master and slave node resources according to the expected job response time.
- Availability considerations may require keeping redundant copies of data blocks, thus doubling or tripling the data volume in HDFS and increasing network traffic.

Sizing a Hadoop cluster can be a complex task. Fujitsu has developed a sizing methodology based on example measurements and a detailed model for all interacting cluster components; a sizing tool for Hadoop helps Fujitsu staff in finding a suitable Hadoop cluster configuration for you.

High Availability

On the software side, high availability features are provided by the Hadoop distribution:

- Disk failures on data nodes expected in Hadoop Clusters and counteracted by data block copies on other data nodes and reassignment of failed MapReduce tasks to other nodes
- Data node failures counteracted in a similar way.
- Name Node high availability problems of earlier Hadoop versions are overcome in different ways
    - In CDH and HDP by an HA concept with active/standby Name Nodes and a quorum of Journal Nodes
    - MapR-FS eliminates the Name Nodes, and MapR's CLDB uses an HA concept with master/slave CLDB instances
- YARN Resource Manager also guarantees HA through standby nodes

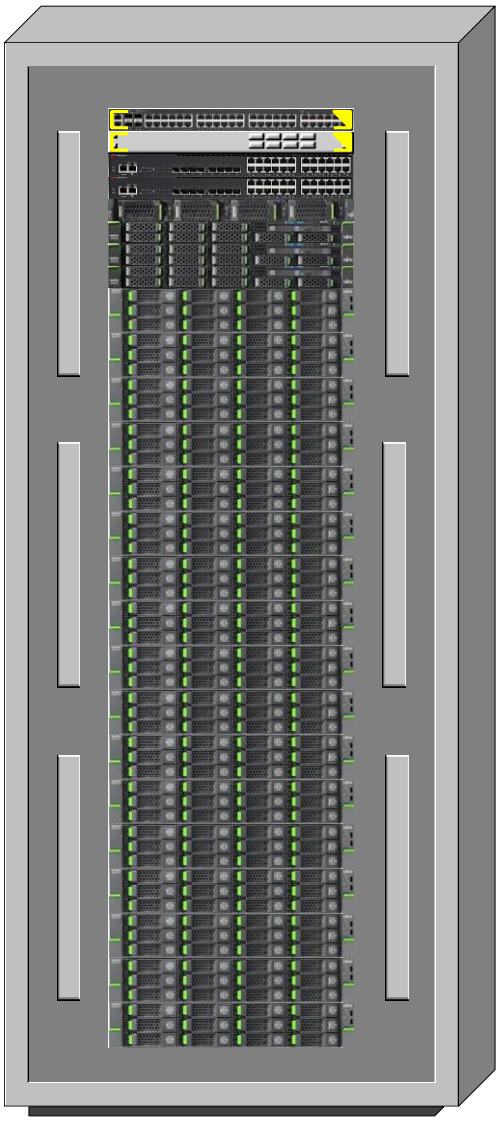
On the hardware side, PRIMEFLEX for Hadoop Entry and Rack variants provides at least RAID-1 for the operating system disks and ECC RAM, rack variants additionally provide redundant/multiple network interfaces and switches.

Appendix

Bill of Material

Example for a fully equipped PRIMEFLEX for Hadoop Rack.

Product no.	Name	Quantity
Rack		1
S26361-K828-V842	PRIMECENTER M2 Rack 42U 1200x800	1
S26361-F5560-E130	PDU basic 16A CEE red 36x10A + 12x16A	2
S26361-F4530-E131	Dummy panel kit 1U plast.	6
S26361-F5560-E2	PDU mounting kit long f PCR M1+ M2	1
T26139-Y1968-E10	Cable powercord rack, 4m, grey	4
Switch 1GbE		1
S26361-K1477-V111	Brocade ICX 6450-48, 48x 1GbE RJ45	1
S26361-F4530-E144	Mounting kit SW2 for BROCADE switches	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
S26361-F1647-E302	Rack installation 1U systems ex factory	1
Switch 10GbE		2
D:ICX7750-48F-M3	ICX7750-48F,48xSFP+,6xQSFP+ports,1xM-SI	2
S26361-F4530-E141	Perforated panel 1U, metal, kit	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
S26361-F1647-E302	Rack installation 1U systems ex factory	1
S26361-F5317-E510	4-Post RMK DX6740T/VDX6740T-1G/ICX7x50	1
D:RPS9-I-E	500W AC power supply with intake airflow	2
D:ICX7750-FAN-I-E	ICX7750 Fan 1x ,port side air exhaust	4
Cables		
S26361-F5317-E40	QSFP+ active Twinax Cable Brocade cust	4
S26361-F3873-E500	SFP+ active Twinax Cable Brocade custom	42
S26361-F3417-E600	Cable CAT 6A, RJ45, customized length	21



Utility Node		1
S26361-K1592-V401	PY RX2530 M4 8x 2.5'	1
S26361-F4051-E110	Intel Xeon Silver 4110 8C 2.10 GHz	2
S26361-F3849-E100	Cooler Kit 2nd CPU	1
S26361-F3694-E10	Independent Mode Installation	2
S26361-F4026-E216	32GB (1x32GB) 2Rx4 DDR4-2666 R ECC	8
S26361-F5675-E240	SSD SATA 6G 240GB Mixed-Use 2.5' H-P EP	8
S26361-F5243-E175	FBU option for PRAID EP4xx	1
S26361-F5243-E12	PRAID EP420i	1
S26361-F5243-E200	TFM module for FBU on PRAID EP420i/e	1
S26361-F3953-E211	PLAN EM 2x 10GB SFP+ OCP interface	1
S26361-F2735-E400	Rack Mount Kit F1-CMA Slim Line	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
S26361-F1647-E302	Rack installation 1U systems ex factory	1
S26361-F2735-E81	Rack Cable Arm 1U	1
S26361-F1452-E100	region kit APAC/EMEA/India	1
S26361-F1790-E243	iRMC advanced pack	1
S26113-F574-E13	Modular PSU 800W platinum hp	2
T26139-Y1968-E100	Cable powercord rack, 4m, black	2
Master Node		2
S26361-K1592-V401	PY RX2530 M4 8x 2.5'	1
S26361-F4051-E234	Intel Xeon Gold 6134 8C 3.20 GHz	2
S26361-F3849-E100	Cooler Kit 2nd CPU	1
S26361-F3694-E10	Independent Mode Installation	2
S26361-F4026-E232	32GB (1x32GB) 2Rx4 DDR4-2666 R ECC	8
S26361-F5675-E240	SSD SATA 6G 240GB Mixed-Use 2.5' H-P EP	6
S26361-F5243-E175	FBU option for PRAID EP4xx	1
S26361-F5243-E12	PRAID EP420i	1
S26361-F5243-E200	TFM module for FBU on PRAID EP420i/e	1
S26361-F3953-E211	PLAN EM 2x 10GB SFP+ OCP interface	1
S26361-F2735-E400	Rack Mount Kit F1-CMA Slim Line	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
S26361-F1647-E302	Rack installation 1U systems ex factory	1
S26361-F2735-E81	Rack Cable Arm 1U	1
S26361-F1452-E100	region kit APAC/EMEA/India	1
S26361-F1790-E243	iRMC advanced pack	1
S26113-F574-E13	Modular PSU 800W platinum hp	2
T26139-Y1968-E100	Cable powercord rack, 4m, black	2
Service Node		1
CZK:1592V301-VGN04	CZK:1592V301-VGN04/Hadoop Service Node	1
S26361-F4051-E110	Intel Xeon Silver 4110 8C 2.10 GHz	2
S26361-F3849-E100	Cooler Kit 2nd CPU	1
S26361-F3694-E10	Independent Mode Installation	2
S26361-F4026-E216	16GB (1x16GB) 1Rx4 DDR4-2666 R ECC	4
S26361-F5550-E130	HD SAS 12G 300GB 10K 512n HOT PL 2.5' EP	2
S26361-F5243-E175	FBU option for PRAID EP4xx	1
S26361-F5243-E12	PRAID EP420i	1
S26361-F5243-E200	TFM module for FBU on PRAID EP420i/e	1
S26361-F3953-E211	PLAN EM 2x 10GB SFP+ OCP interface	1
S26361-F2735-E400	Rack Mount Kit F1-CMA Slim Line	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
S26361-F2735-E81	Rack Cable Arm 1U	1
S26361-F1452-E100	region kit APAC/EMEA/India	1

S26361-F1790-E243	iRMC advanced pack	1
S26113-F575-E13	Modular PSU 450W platinum hp	2
T26139-Y1968-E100	Cable powercord rack, 4m, black	2
Datanodes		17
S26361-K1567-V116	PY RX2540 M4 12x 3.5' +4 @1Ctrl	1
S26361-F4051-E110	Intel Xeon Silver 4110 8C 2.10 GHz	2
S26361-F3849-E100	Cooler Kit 2nd CPU	1
S26361-F3694-E10	Independent Mode Installation	2
S26361-F4026-E216	16GB (1x16GB) 1Rx4 DDR4-2666 R ECC	8
S26361-F3853-E30	4x REAR 2.5' SAS/SATA HDD/SSD	1
S26361-F5635-E400	HD SAS 12G 4TB 7.2K 512e HOT PL 3.5' BC	12
S26361-F5675-E240	SSD SATA 6G 240GB Mixed-Use 2.5' H-P EP	4
S26361-F4042-E204	PRAID EP540i LP	1
S26361-F3953-E211	PLAN EM 2x 10GB SFP+ OCP interface	1
S26361-F2735-E175	Rack Mount Kit F1 CMA QRL LV	1
S26361-F4530-E10	Mounting of RMK in symmetrical racks	1
SNP:SY-F1647E301-P	Rack installation ex works	1
S26361-F1452-E100	region kit APAC/EMEA/India	1
S26361-F1790-E243	iRMC advanced pack	1
S26113-F574-E13	Modular PSU 800W platinum hp	2
T26139-Y1968-E100	Cable powercord rack, 4m, black	2

Contact

FUJITSU
Fujitsu Technology Solutions GmbH
Mies-van-der-Rohe-Str. 8, 80807 Munich, Germany
E-mail: cic@ts.fujitsu.com
Website: www.fujitsu.com

2018-03-15 CE EN