Topic-1 Hadoop

Big Data

- Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools.

- Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set.

Big Data

- In other words, the data set has grown so large that it is difficult to manage and even harder to garner value out of it. The primary difficulties are the
  - acquisition,
  - storage,
  - searching,
  - sharing,
  - analytics, and
  - visualization of data.

The Arrival Of Analytics

- As analytics and research were applied to large data sets, scientists came to the conclusion that more is better—in this case, more data, more analysis, and more results.
- Researchers started to incorporate related data sets, unstructured data, archival data, and real-time data into the process, which in turn gave birth to what we now call Big Data.

Big Data

- In the business world, Big Data is all about ***opportunity***. According to IBM, every day we create 2.5 quintillion (2.5×1018) bytes of data, so much that 90 percent of the data in the world today has been created in the last two years.
- These data come from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals, to name just a few.
- That is the catalyst for Big Data, along with the more important fact that all of these data have intrinsic value that can be extrapolated using analytics, algorithms, and other techniques.

Where Is The Value?

- Extracting value is much more easily said than done. Big Data is full of challenges, ranging from the technical to the conceptual to the operational, any of which can derail the ability to discover value and leverage what Big Data is all about.

4 V's

- 1. **Volume**. Big Data comes in one ***size***: large. Enterprises are awash with data, easily amassing terabytes and even petabytes of information.
- 2. **Variety**. Big Data extends beyond structured data to include unstructured data of all ***varieties***: text, audio, video, click streams, log files, and more.
- 3. **Veracity**. The massive amounts of data collected for Big Data purposes can lead to statistical errors and misinterpretation of the collected information. ***Purity*** of the information is critical for value.
- 4. **Velocity**. Often ***time sensitive***, Big Data must be used as it is ***streamin***g into the enterprise in order to maximize its value to the business, but it must also still be available from the ***archival*** sources as well

4 V's and processes

- The complexity of Big Data does not end with just four dimensions.
- There are other factors at work as well: the processes that Big Data drives.
- These processes are a conglomeration of technologies and analytics that are used to define the value of data sources, which ***translates to actionable elements that move businesses forward***.

Technologies

- Many of those technologies or concepts are not new but have come to fall under the umbrella of Big Data.
  - **Business Intelligence**
    - This consists of a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data.
    - BI delivers actionable information, which helps enterprise users make better business decisions using fact-based support systems.
    - BI works by using an in-depth analysis of detailed business data, provided by databases, application data, and other tangible data sources.
    - In some circles, BI can provide historical, current, and predictive views of business operations

Technologies

- **Data Mining**
  - This is a process in which data are analyzed from different perspectives and then turned into summary data that are deemed useful.
  - Data mining is normally used with data at rest or with archival data.
  - Data mining techniques focus on modeling and knowledge discovery for predictive, rather than purely descriptive, purposes—an ideal process for uncovering new patterns from large data sets.

Technologies

- **Statistical applications**.
  - These look at data using algorithms based on statistical principles and normally concentrate on data sets related to polls, census, and other static data sets.
  - Statistical applications ideally deliver sample observations that can be used to study populated data sets for the purpose of estimating, testing, and predictive analysis.
  - Empirical data, such as surveys and experimental reporting, are the primary sources for analyzable information

Technologies

- **Predictive analysis.**
  - This is a subset of statistical applications in which data sets are examined to come up with predictions, based on trends and information gleaned from databases.
  - Predictive analysis tends to be big in the financial and scientific worlds, where trending tends to drive predictions, once external elements are added to the data set.
  - One of the main goals of predictive analysis is to identify the risks and opportunities for business process, markets, and manufacturing

Technologies

- **Data modeling.**
  - This is a conceptual application of analytics in which multiple "what-if" scenarios can be applied via algorithms to multiple data sets.
  - Ideally, the modeled information changes based on the information made available to the algorithms, which then provide insight to the effects of the change on the data sets.
  - Data modeling works hand in hand with data visualization, in which uncovering information can help with a particular business endeavor.

Technologies

- The preceding analysis categories constitute only a portion of where Big Data is headed and why it has intrinsic value to business.
- That value is driven by the never-ending quest for a competitive advantage, encouraging organizations to turn to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to ***help them decide on their next move***.
- Big data encompasses everything from dollar transactions to tweets to images to audio.
- Therefore, taking advantage of big data requires that all this ***information be integrated for analysis and data management***

Varying data structures

- **Structured data** is characterized by a high degree of organization and is typically the kind of data you see in relational databases or spreadsheets.
- Because of its defined structure, it maps easily to one of the standard data types (or user-defined types that are based on those standard types).
- It can be searched using standard search algorithms and manipulated in well-defined ways

Varying data structures

- **Structured data sources**
  - Sensor Data (RFID)
  - Point of Sale data
  - Financial Data
  - Input Data
  - Click Stream Data

Varying data structures

- **Semistructured data** (such as what you might see in log files) is a bit more difficult to understand than structured data. Normally, this kind of data is stored in the form of text files, where there is some degree of order — for example, tab delimited files, where columns are separated by a tab character.
- So instead of being able to issue a database query for a certain column and knowing exactly what you're getting back, users typically need to explicitly assign data types to any data elements extracted from semi-structured data sets.

Varying data structures

- **Semistructured Data sources**
  - Semi-structured data does not necessarily conform to a fixed schema (that is, structure) but may be self-describing and may have simple label/value pairs. For example, label/value pairs might include: <family>=Jones, <mother>=Jane, and <daughter>=Sarah.
  - XML, JSON

Varying data structures

- **Unstructured data** has none of the advantages of having structure coded into a data set
- Its analysis by way of more traditional approaches is difficult and costly at best, and logistically impossible at worst.
- Just imagine having many years' worth of notes typed by call center operators that describe customer observations.
- Without a robust set of text analytics tools, it would be extremely tedious to determine any interesting behavior patterns.
- Moreover, the sheer volume of data in many cases poses virtually insurmountable challenges to traditional data mining techniques, which, even when conditions are good, can handle only a fraction of the valuable data that's available

Varying data structures

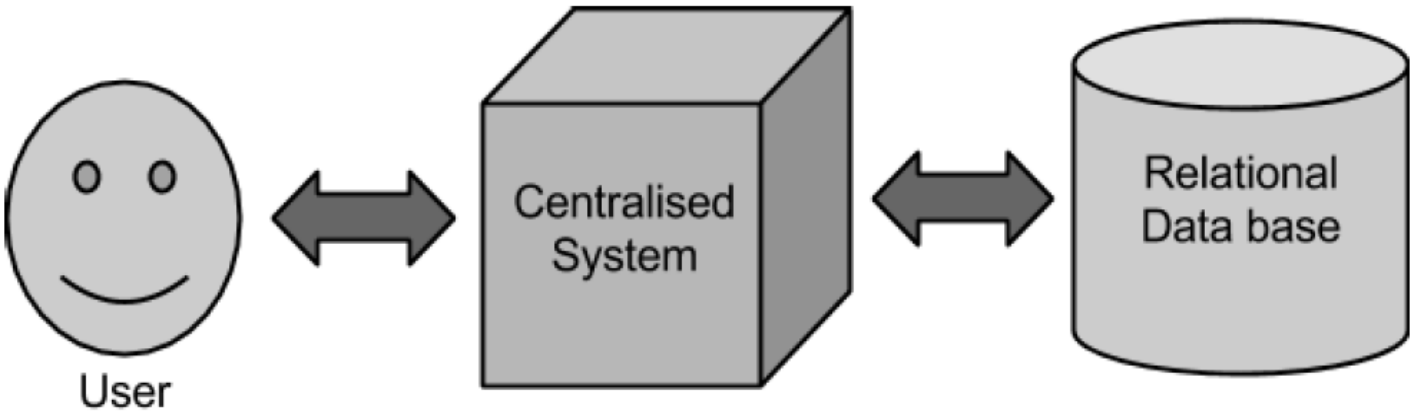
- **<u>Unstructured Data sources</u>**
  - Satellite Images
  - Photographs and Videos
  - Radar and Sonar Data
  - Text Internal to your Company
  - Social Media Data
  - Mobile Data (Messages, Location etc)
  - Website content (Youtube, Instagrametc)

A playground for Data Scientists

- A *data scientist is a computer scientist who loves data (lots of data) and the* sublime challenge of figuring out ways to squeeze every drop of value out of that abundant data.
- A *data playground is an enterprise store of many terabytes* (or even petabytes) of data that data scientists can use to develop, test, and enhance their analytical "toys."

Traditional Enterprise Approach

- In this approach, an enterprise will have a **computer to store and process big data**. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

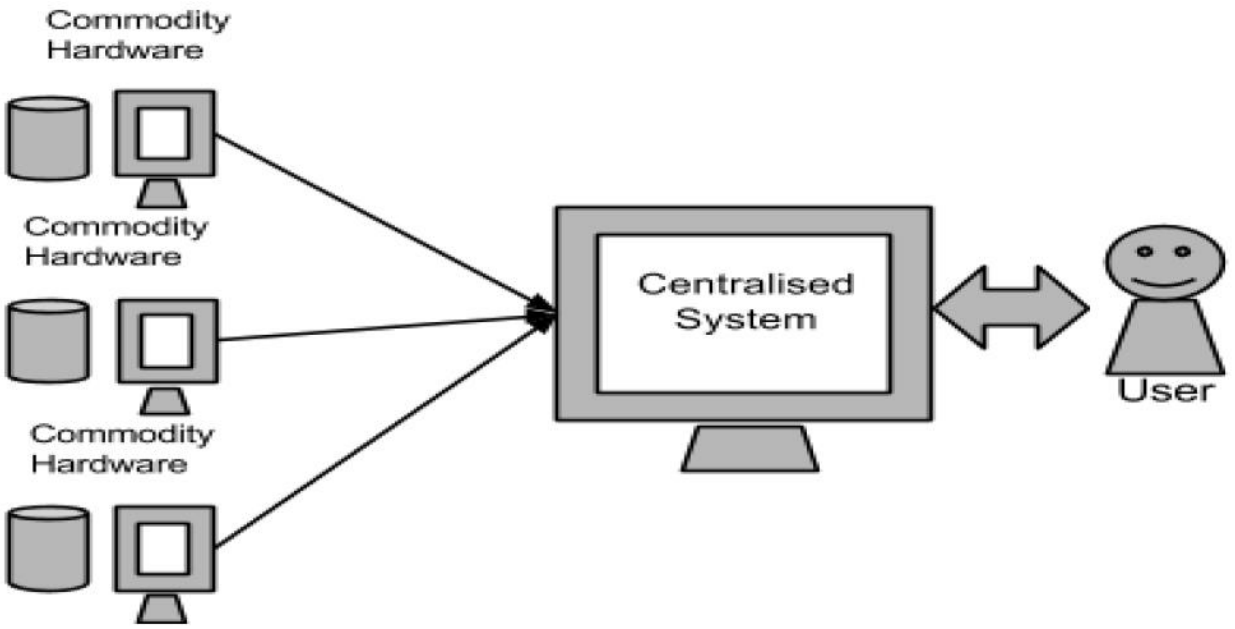


Traditional Enterprise Approach

- **Limitation**
- This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But when it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck

Google's Solution

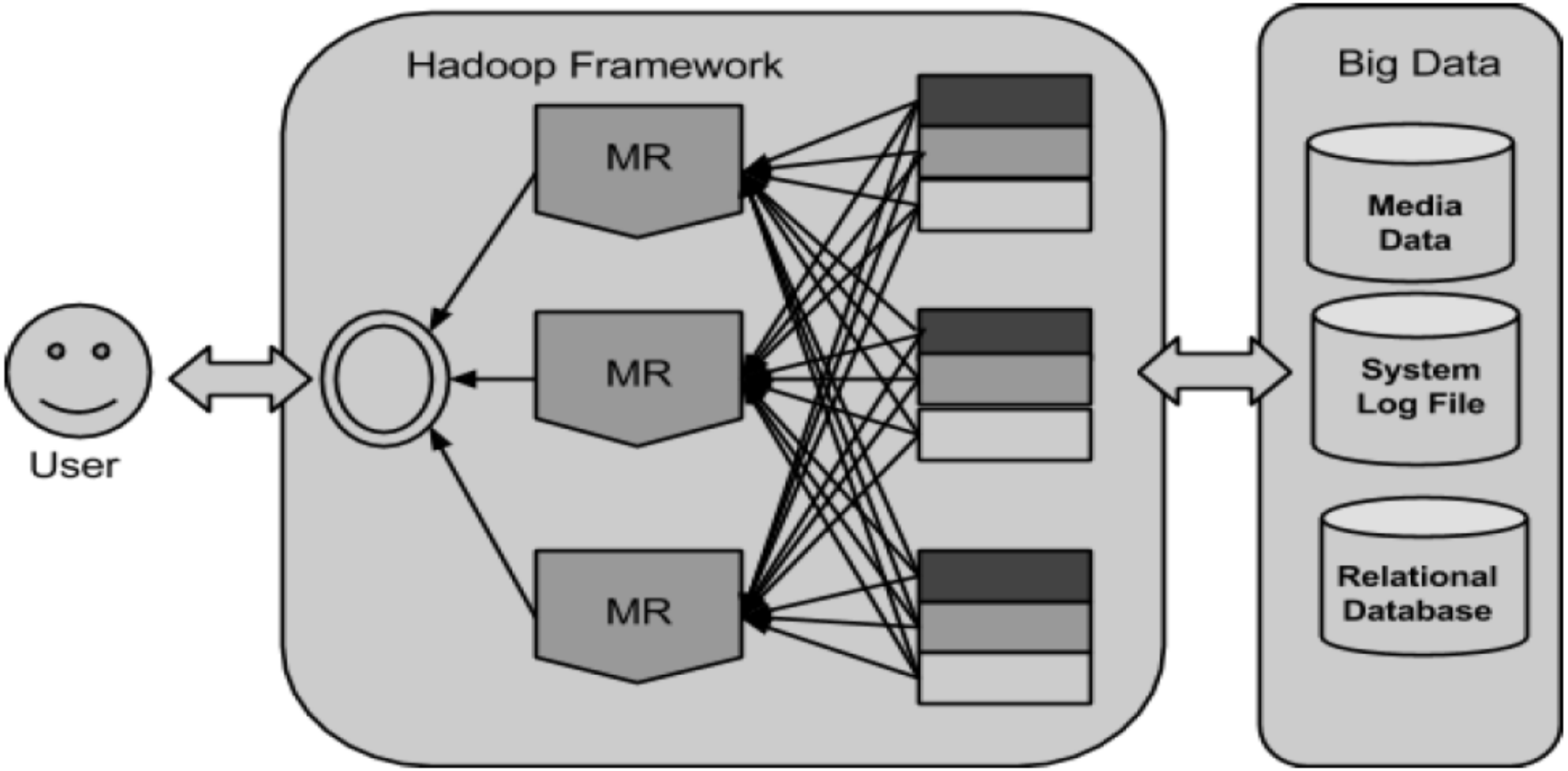
- Google solved this problem using an algorithm called **MapReduce**. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.



Hadoop

- Using the solution provided by Google, **Doug Cutting and his team developed an Open Source Project called HADOOP.**
- Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, *Hadoop is used to develop applications* that could perform complete statistical analysis on huge amounts of data.

Hadoop Framework



The Origin and Design of Hadoop

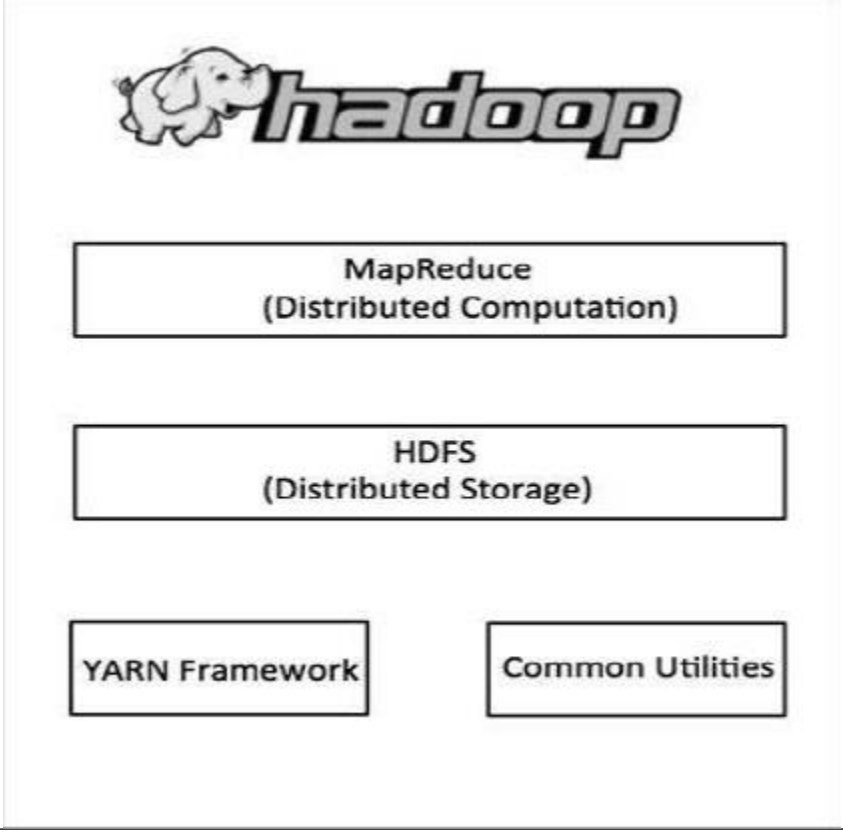
- At its core, Hadoop is a framework for storing data on large ***clusters of commodity hardware (***everyday computer hardware that is affordable and easily available) and running applications against that data.
- A cluster is a group of interconnected computers (known as nodes) that can work together on the same problem. Using networks of affordable compute resources to acquire business insight is the key value proposition of Hadoop.

The Origin and Design of Hadoop

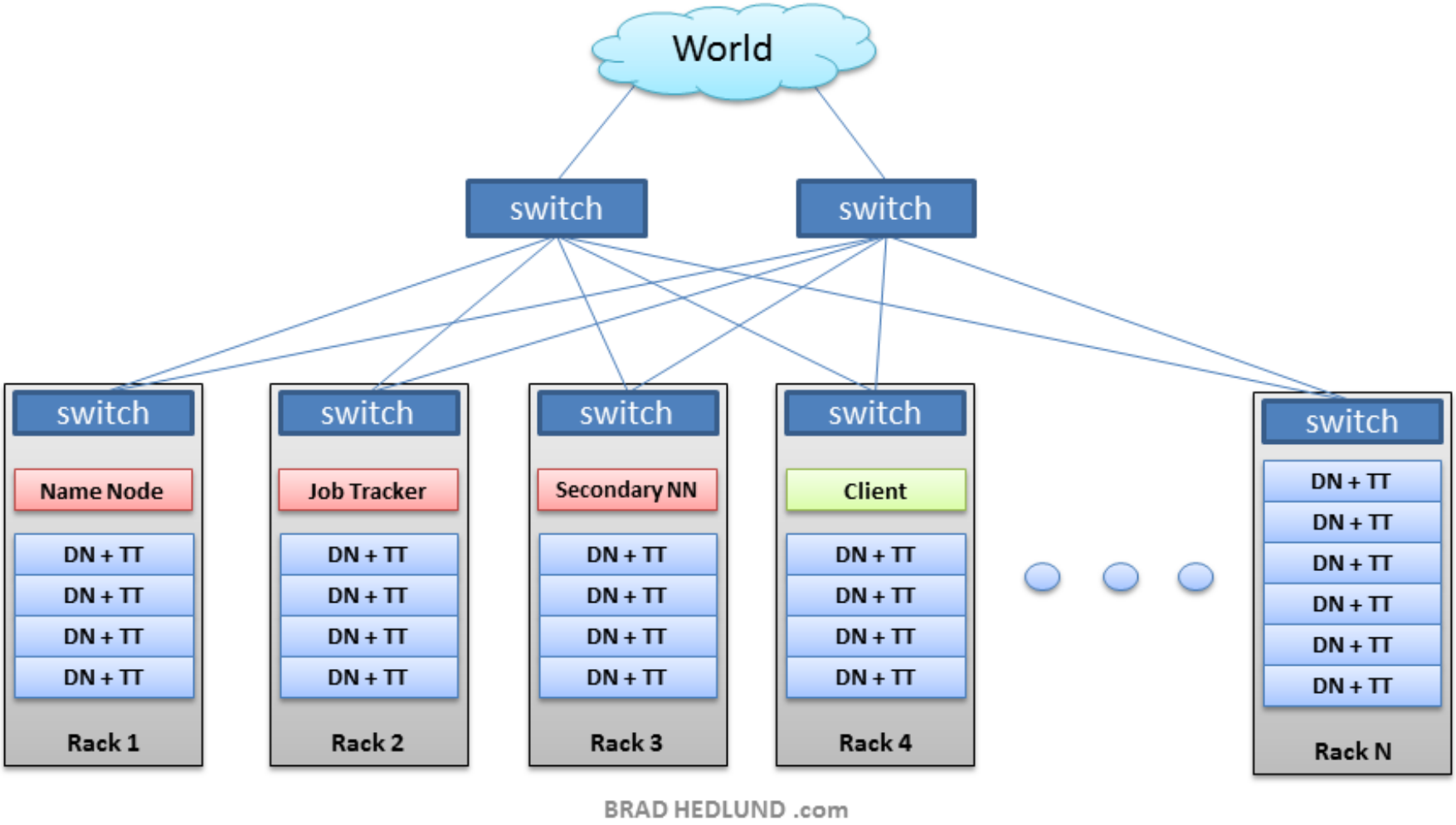
- As for that name, Hadoop, don't look for any major significance there; it's simply the name that Doug Cutting's son gave to his stuffed elephant. (Doug Cutting is, of course, the co-creator of Hadoop.)

Hadoop Architecture

- At its core, Hadoop has two major layers namely:
  - (a) Processing/Computation layer (MapReduce), and
  - (b) Storage layer (Hadoop Distributed File System).



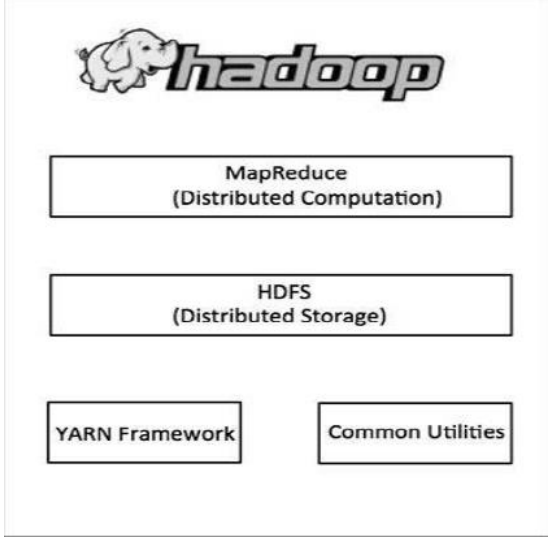
Hadoop Cluster



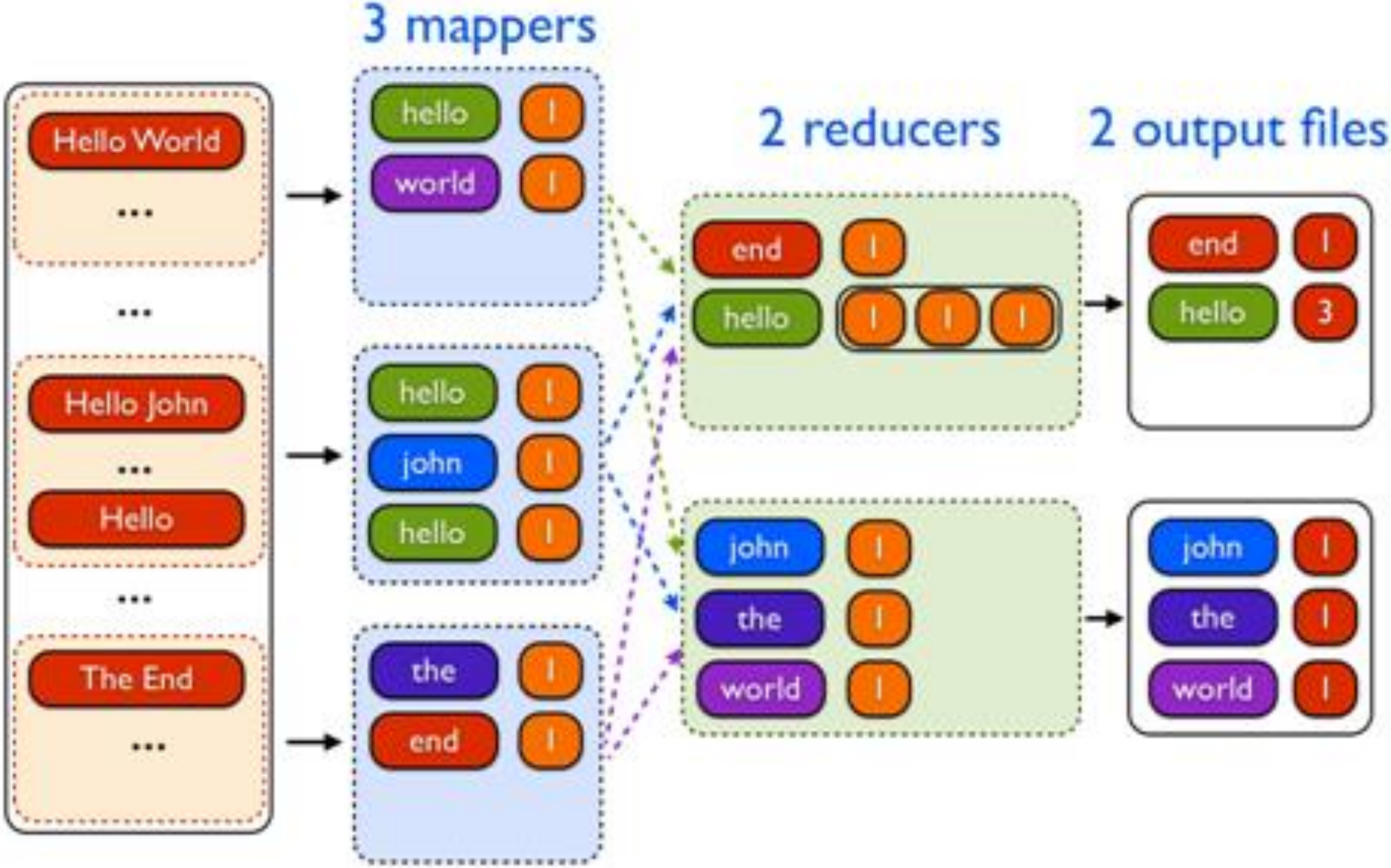
BRAD HEDLUND .com

MapReduce

- MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.
- The MapReduce program runs on Hadoop which is an Apache open-source framework.



Hadoop MapReduce



Hadoop Distributed File System

- The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.
- Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:
  - **Hadoop Common Utilities:** These are Java libraries and utilities required by other Hadoop modules.
  - **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.

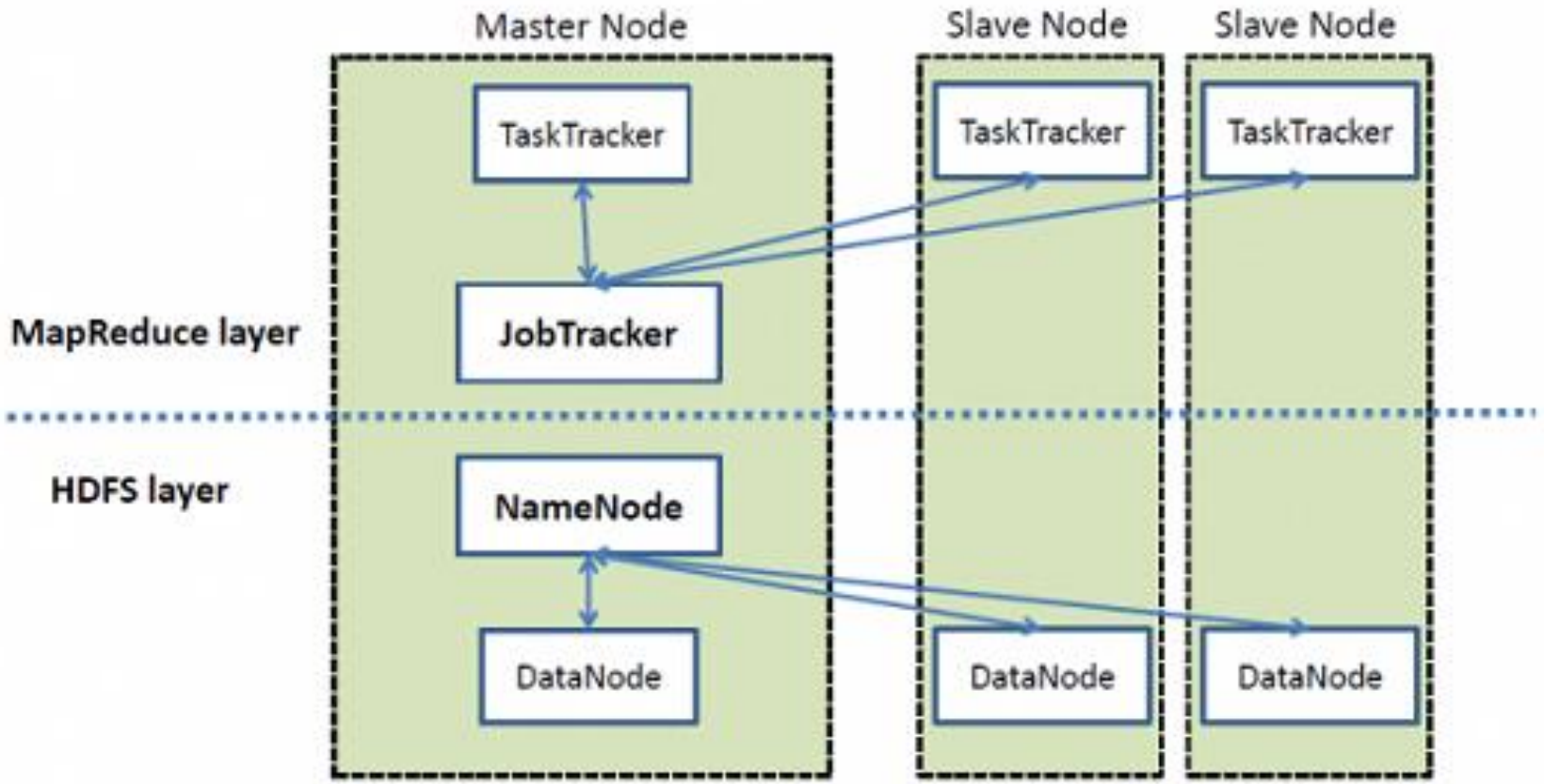
The Origin and Design of Hadoop

- Hadoop consists of two main components: a distributed processing framework named MapReduce (which is now supported by a component calledYARN) and a distributed file system known as the Hadoop distributed file system, or HDFS.
- An application that is running on Hadoop gets its work divided among the nodes (machines) in the cluster, and HDFS stores the data that will be processed.
- A Hadoop cluster can span thousands of machines, where HDFS stores data, and MapReduce jobs do their processing near the data, which keeps I/O costs low.
- MapReduce is extremely flexible, and enables the development of a wide variety of applications.

The Origin and Design of Hadoop

- A Hadoop cluster is a form of *compute cluster, a* type of cluster that's used mainly for computational purposes. In a compute cluster, many computers (*compute nodes) can share computational workloads* and take advantage of a very large aggregate bandwidth across the cluster.
- Hadoop clusters typically consist of a few *master nodes, which control the* storage and processing systems in Hadoop, and many *slave nodes, which store* all the cluster's data and is also where the data gets processed.

High Level Architecture of Hadoop



Distributed processing with MapReduce

- MapReduce involves the processing of a sequence of operations on distributed data sets. The data consists of key-value pairs, and the computations have only two phases: a map phase and a reduce phase. User-defined MapReduce jobs run on the compute nodes in the cluster

How Does Hadoop Work?

- It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput.
- Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

How Does Hadoop Work?

- Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs:
  - Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
  - These files are then distributed across various cluster nodes for further processing.
  - HDFS, being on top of the local file system, supervises the processing.
  - Blocks are replicated for handling hardware failure.
  - Checking that the code was executed successfully.
  - Performing the sort that takes place between the map and reduce stages.
  - Sending the sorted data to a certain computer.
  - Writing the debugging logs for each job.

Advantagesof Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

MapReduce

Generally speaking, a MapReduce job runs as follows:

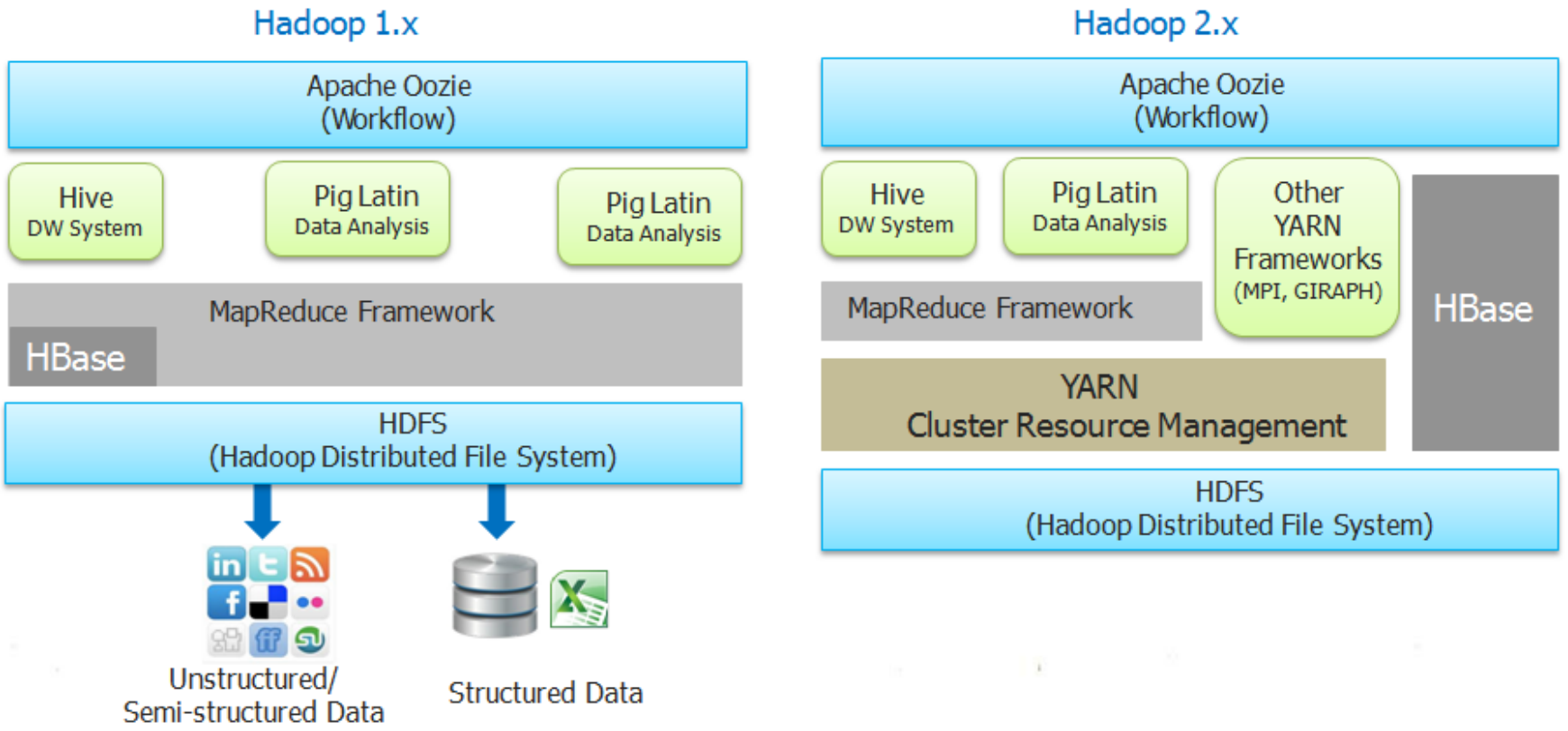
1. During the Map phase, input data is split into a large number of fragments, each of which is assigned to a map task.
2. These map tasks are distributed across the cluster.
3. Each map task processes the key-value pairs from its assigned fragment and produces a set of intermediate key-value pairs.
4. The intermediate data set is sorted by key, and the sorted data is partitioned into a number of fragments that matches the number of reduce tasks.
5. During the Reduce phase, each reduce task processes the data fragment that was assigned to it and produces an output key-value pair.
6. These reduce tasks are also distributed across the cluster and write their output to HDFS when finished.

MapReduce

- The Hadoop MapReduce framework in earlier (pre-version 2) Hadoop releases has a single master service called a JobTracker and several slave services called TaskTrackers, one per node in the cluster.
- When you submit a MapReduce job to the JobTracker, the job is placed into a queue and then runs according to the scheduling rules. As you might expect, the JobTracker manages the assignment of map-and-reduce tasks to the TaskTrackers.

YARN

- With Hadoop 2, a new resource management system is in place called YARN (short for *Yet Another Resource Manager). YARN provides generic scheduling* and resource management services so that you can run more than just Map Reduce applications on your Hadoop cluster. The JobTracker / TaskTracker architecture could only run MapReduce.

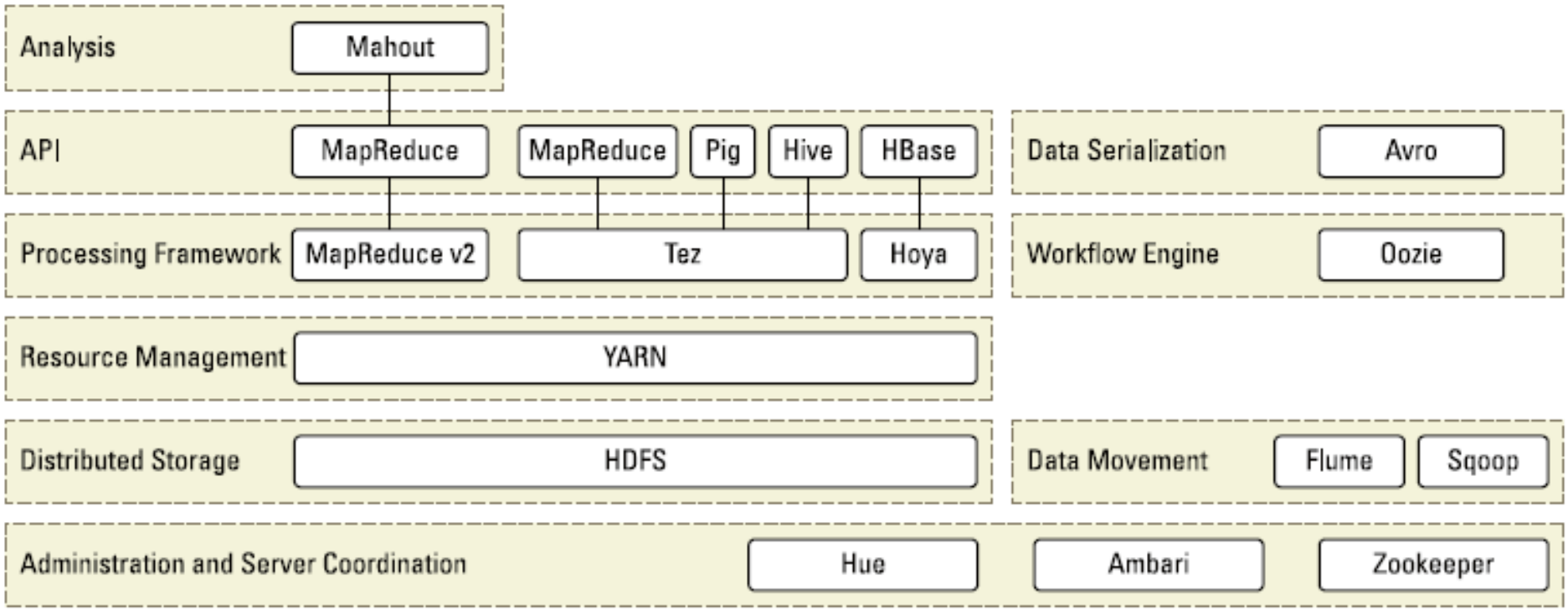

Hadoop 1.x
Apache Oozie
(Workflow)
Hive
DW System
Pig Latin
Data Analysis
Pig Latin
Data Analysis
MapReduce Framework
HBase
HDFS
(Hadoop Distributed File System)
Unstructured/
Semi-structured Data
Structured Data
Hadoop 2.x
Apache Oozie
(Workflow)
Hive
DW System
Pig Latin
Data Analysis
Other
YARN
Frameworks
(MPI, GIRAPH)
HBase
MapReduce Framework
YARN
Cluster Resource Management
HDFS
(Hadoop Distributed File System)

HDFS

- HDFS also has a master/slave architecture:
  - **Master service:** Called a *NameNode*, it controls access to data files.
  - **Slave services:** Called *DataNodes*, they're distributed one per node in the cluster. DataNodes manage the storage that's associated with the nodes on which they run, serving client read and write requests, among other tasks.

Apache Hadoop ecosystem

- Other open source components that are typically seen in a Hadoop deployment. Hadoop is more than MapReduce and HDFS: It's also a family of related projects (an ecosystem, really) for distributed computing and large-scale data processing.


Analysis
Mahout
API
MapReduce
MapReduce
Pig
Hive
HBase
Data Serialization
Avro
Processing Framework
MapReduce v2
Tez
Hoya
Workflow Engine
Oozie
Resource Management
YARN
Distributed Storage
HDFS
Data Movement
Flume
Sqoop
Administration and Server Coordination
Hue
Ambari
Zookeeper

Project Name	***Description***
Ambari	An integrated set of Hadoop administration tools for installing, monitoring, and maintaining a Hadoop cluster. Also included are tools to add or remove slave nodes.
Avro	A framework for the efficient *serialization* (a kind of transformation) of data into a compact binary format
Flume	A data flow service for the movement of large volumes of log data into Hadoop
HBase	A distributed columnar database that uses HDFS for its underlying storage. With HBase, you can store data in extremely large tables with variable column structures
HCatalog	A service for providing a relational view of data stored in Hadoop, including a standard approach for tabular data

Project Name	Description
Hive	A distributed data warehouse for data that is stored in HDFS; also provides a query language that's based on SQL (HiveQL)
Hue	A Hadoop administration interface with handy GUI tools for browsing files, issuing Hive and Pig queries, and developing Oozie workflows
Mahout	A library of machine learning statistical algorithms that were implemented in MapReduce and can run natively on Hadoop
Oozie	A workflow management tool that can handle the scheduling and chaining together of Hadoop applications
Pig	A platform for the analysis of very large data sets that runs on HDFS and with an infrastructure layer consisting of a compiler that produces sequences of MapReduce programs and a language layer consisting of the query language named Pig Latin
Sqoop	A tool for efficiently moving large amounts of data between relational databases and HDFS
ZooKeeper	A simple interface to the centralized coordination of services (such as naming, configuration, and synchronization) used by distributed applications

Data Warehouse and Hadoop

DW- Hadoop

- Agriculture:
  - Data warehousing: Cost of farm production and optimization, yield analysis, agricultural goods commodity pricing / trading analysis.
  - Hadoop / Internet of Things Hadoop: Analysis and optimization of plowing patterns, fertilization, readiness for harvesting, and moisture content (from sensors in the field and weather data).

DW- Hadoop

- Automotive Manufacturing:
  - Data warehousing: Cost and quality of manufacturing, supply chain analysis, warranty analysis, sales and marketing analysis, human capital management.
  - Hadoop / Internet of Things: Analysis of customer sentiment and analysis of connected vehicles including component failure, need for service and service scheduling, driving history (and automated car), driver emergency detection and response.

DW- Hadoop

- Banking:
  - Data warehousing: Single view of customer across financial offering channels, financial analysis, fraud detection, credit worthiness, human resource management, real estate management and optimization.
  - Hadoop / Internet of Things: Fraud detection, risk analysis, and customer sentiment.

DW- Hadoop

- Communications:
  - Data warehousing: Pricing strategies and finances, customer support and service, marketing analysis, supply chain, logistics and process optimization, regulatory compliance, real estate optimization, and human capital management.
  - Hadoop / Internet of Things: Analysis of social data, mobile device usage, network quality and availability (using sensors), network fraud detection, and for Internet of Things, extended network management and optimization.

DW- Hadoop

- Consumer Packaged Goods:
  - Data warehousing: Analysis of sales, marketing, suppliers, manufacturing, logistics, consumer trends, and risk.
  - Hadoop / Internet of Things: Analysis of promotional effectiveness (through social media and in-store sensors), supply chain, state of manufactured goods during transport, product placement in retail, and risk.

DW- Hadoop

- Education and Research:
  - Data warehousing: Financial analysis of institution or facility, staffing and human capital management, and alumni profiling and donation patterns.
  - Hadoop / Internet of Things: Analysis of students at risk (using sensor data), research data from sensors, and facilities monitoring and utilization optimization.

DW- Hadoop

- Healthcare Payers:
  - Data warehousing: Analysis of cost of care, quality of care, risk, and fraud.
  - Hadoop / Internet of Things: Analysis of sentiment of insured customers, risk, and fraud.

DW- Hadoop

- Healthcare Providers:
  - Data warehousing: Analysis of cost of care, quality of care analysis, staffing and human resources, and risk.
  - Hadoop / Internet of Things: Disease and epidemic pattern research, patient monitoring, facilities monitoring and optimization, patient sentiment analysis, and risk analysis.

DW- Hadoop

- High Tech and Industrial Manufacturing:
  - Data warehousing: Supplier and distributor analysis, logistics management, quality of manufacturing and warranty analysis.
  - Hadoop / Internet of Things: Shop-floor production and quality analysis, quality of sub-assembly analysis, product failure and pending failure analysis, and automated service requests.

DW- Hadoop

- Insurance (Property and Casualty):
  - Data warehousing: Sales and marketing analysis, human resources analysis, and risk analysis.
  - Hadoop / Internet of Things: Customer sentiment analysis and risk analysis.

DW- Hadoop

- Law Enforcement:
  - Data warehousing: Logistics optimization, crime statistics analysis, and human resources optimization.
  - Hadoop / Internet of Things: Threat analysis (from social media and video capture identification).

DW- Hadoop

- Media and Entertainment:
  - Data warehousing: Analysis of viewer preferences, media channel popularity, advertising sales, and marketing promotions.
  - Hadoop / Internet of Things: Viewing habit analysis (from set-top boxes), analysis of customer behavior at entertainment venues, and customer sentiment analysis.

DW- Hadoop

- Oil and Gas:
  - Data warehousing: Analysis of drilling exploration costs, potential exploration sites, production, human resources, and logistics optimization
  - Hadoop / Internet of Things: Drilling exploration sensor analysis (failure prevention)

DW- Hadoop

- Pharmaceuticals:
  - Data warehousing: Clinical trials including drug interaction research, test subject outcome analysis, research and production financial analysis, sales and marketing analysis, and human resources analysis.
  - Hadoop / Internet of Things: Analysis of clinical research data from sensors, social behavior and disease tracking (from social media), and genomics research.

DW- Hadoop

- Retail:
  - Data warehousing: Market basket analysis, sales analysis, supply chain optimization, real estate optimization, and logistics and distribution optimization.
  - Hadoop / Internet of Things: Omni-channel analysis and customer sentiment analysis.

DW- Hadoop

- Transportation and Logistics:
  - Data warehousing: Equipment and crew logistics and routing, sales and marketing analysis, real estate optimization, and human resources analysis and optimization.
  - Hadoop / Internet of Things: Traffic optimization (from highway sensor data), traffic safety analysis and control, equipment performance and potential failure analysis (from on-board sensors), logistics management (from sensors), and customer sentiment analysis.

DW- Hadoop

- Utilities:
  - Data warehousing: Logistics optimization, grid power delivery analysis and optimization, customer energy utilization, and human resources analysis and optimization.
  - Hadoop / Internet of Things: Analysis of data from smart meters for grid optimization and status, pro-active maintenance optimization.

Examining the Various Hadoop Offerings

- Hadoop is available from either the Apache Software Foundation or from companies that offer their own Hadoop distributions.
- Only products that are available directly from the Apache Software Foundation can be called Hadoop releases.
- Products from other companies can include the official Apache Hadoop release files, but products that are "forked" from (and represent modified or extended versions of) the Apache Hadoop source tree are not supported by the Apache Software Foundation

Examining the Various Hadoop Offerings

- Apache Hadoop Release Notes

Comparing distributions

- Red Hat is, for many people, the model of how to successfully make money in the open source software market.
- What Red Hat has done is to take Linux (an open source operating system), bundle all its required components, build a simple installer, and provide paid support to any customers.
- In the same way that Red Hat has provided a handy packaging for Linux, a number of companies have bundled Hadoop and some related technologies into their own Hadoop distributions

Hadoop Distributions

1. **Cloudera**
2. **EMC**
3. **Hortonworks**
4. **IBM**
5. **Intel**
6. **MapR**

Cloudera (www.cloudera.com/)

- Perhaps the best-known player in the field.
- Cloudera is able to claim Doug Cutting, Hadoop's co-founder, as its chief architect.
- **Cloudera Enterprise,**includes the Cloudera Distribution for Hadoop (CDH), an open-source-based distribution of Hadoop and its related projects as well as its proprietary Cloudera Manager. Also included is a technical support subscription for the core components of CDH.

Cloudera

- Cloudera's primary business model has long been based on its ability to leverage its popular CDH distribution and provide paid services and support.
- In the fall of 2013, Cloudera formally announced that it is focusing on adding proprietary value-added components on top of open source Hadoop to act as a differentiator.
- Also, Cloudera has made it a common practice to accelerate the adoption of alpha- and beta-level open source code for the newer Hadoop releases. Its approach is to take components it consider's to be mature and retrofit (add (a component) to something that did not have it when manufactured.) them into the existing production ready open source libraries that are included in its distribution.

EMC (www.gopivotal.com)

- Pivotal HD, the Apache Hadoop distribution from EMC, natively integrates EMC's massively parallel processing (MPP) database technology (formerly known as Greenplum, and now known as HAWQ) with Apache Hadoop. The result is a high-performance Hadoop distribution with true SQL processing for Hadoop. SQL-based queries and other business intelligence tools can be used to analyze data that is stored in HDFS.

Hortonworks (www.hortonworks.com)

- Another major player in the Hadoop market,.
- Hortonworks has the largest number of committers (Committers are the gatekeepers of Apache projects and have the power to approve code changes.) and code contributors for the Hadoop ecosystem components.
- The Hortonworks business model is based on its ability to leverage its popular HDP distribution and provide paid services and support.
- However, it does not sell proprietary software.

Hortonworks (www.hortonworks.com)

- Hortonworks has forged a number of relationships with established companies in the data management industry: Teradata, Microsoft, Informatica, and SAS, for example. Though these companies don't have their own, in-house Hadoop offerings, they collaborate with Hortonworks to provide integrated Hadoop solutions with their own product sets.
- The Hortonworks Hadoop offering is the Hortonworks Data Platform (HDP), which includes Hadoop as well as related tooling and projects. Also unlike Cloudera, Hortonworks releases only HDP versions with production-level code from the open source community.

IBM (www.ibm.com/software/data/infosphere/biginsights)

- Big Blue offers a range of Hadoop offerings, with the focus around value added on top of the open source Hadoop stack:
  - ***InfoSphere BigInsights***:*This software-based offering includes a* number of Apache Hadoop ecosystem projects, along with additional software to provide additional capability. The focus of InfoSphere BigInsights is on making Hadoop more readily consumable for businesses. As such, the proprietary enhancements are focused on standards-based SQL support, data security and governance, spreadsheet-style analysis for business users, text analytics, workload management, and the application development life cycle.
  - ***PureData System for Hadoop***:*This hardware- and software-based* appliance is designed to reduce complexity, the time it takes to start analyzing data, as well as IT costs. It integrates InfoSphere BigInsights (Hadoop-based software), hardware, and storage into a single, easy-to-manage system.

Intel (hadoop.intel.com)

- The Intel Distribution for Apache Hadoop (Intel Distribution) provides distributed processing and data management for enterprise applications that analyze big data.
- Key features include excellent performance with optimizations for Intel Xeon processors, Intel SSD storage, and Intel 10GbE networking; data security via encryption and decryption in HDFS, and role-based access control with cell-level granularity in HBase (you can control who's allowed to see what data down to the cell level, in other words); improved Hive query performance; support for statistical analysis with a connector for R, the popular open source statistical package; and analytical graphics through Intel Graph Builder.

Intel (hadoop.intel.com)

- The motivations for Intel are simple, though: Hadoop is a strategic platform, and it will require significant hardware investment, especially for larger deployments.
- Though much of the initial discussion around hardware reference architectures for Hadoop — the recommended patterns for deploying hardware for Hadoop clusters — have focused on commodity hardware, increasingly we are seeing use cases where more expensive hardware can provide significantly better value.
- It's with this situation in mind that Intel is keenly interested in Hadoop. It's in Intel's best interest to ensure that Hadoop is optimized for Intel hardware, on both the higher end and commodity line.

Intel (hadoop.intel.com)

- The Intel Distribution comes with a management console designed to simplify the configuration, monitoring, tuning, and security of Hadoop deployments. This console includes automated configuration with Intel Active Tuner; simplified cluster management; comprehensive system monitoring and logging; and systematic health checking across clusters.

MapR (www.mapr.com)

- For a complete distribution for Apache Hadoop and related projects that's independent of the Apache Software Foundation, look no further than MapR.
- Boasting no Java dependencies or reliance on the Linux file system, MapR is being promoted as the only Hadoop distribution that provides full data protection, no single points of failure, and significant ease-of-use advantages.
- Three MapR editions are available: M3, M5, and M7.
- The M3 Edition is free and available for unlimited production use; MapR M5 is an intermediate-level subscription software offering; and MapR M7 is a complete distribution for Apache Hadoop and HBase that includes Pig, Hive, Sqoop, and much more.

MapR

- The MapR distribution for Hadoop is most well-known for its file system, which has a number of enhancements not included in HDFS, such as NFS access and POSIX compliance (long story short, this means you can mount the MapR file system like it's any other storage device in your Linux instance and interact with data stored in it with any standard file applications or commands), storage volumes for specialized management of data policies, and advanced data replication tools. MapR also ships a specialized version of HBase, which claims higher reliability, security, and performance than Apache HBase.

Working with in-database MapReduce

- When MapReduce processing occurs on structured data in a relational database, the process is referred to as *in-database MapReduce*.
- *One implementation* of a hybrid technology that combines MapReduce and relational databases for the analysis of analytical workloads is **HadoopDB**, a research project that originated a few years ago at Yale University. HadoopDB was designed to be a free, highly scalable, open source, parallel database management system. Tests at Yale showed that HadoopDB could achieve the performance of parallel databases, but with the scalability, fault tolerance, and flexibility of Hadoop-based systems.
- More recently, Oracle has developed an in-database Hadoop prototype that makes it possible to run Hadoop programs written in Java naturally from SQL. Users with an existing database infrastructure can avoid setting up a Hadoop cluster and can execute Hadoop jobs within their relational databases.