ACID test

A test applied to data for atomicity, consistency, isolation, and durability

Acunu

Acunu is a new low-level data storage layer that replaces the traditional file system, though its initial target is Cassandra rather than Hadoop. By writing a kernel-level key/value store called Castle, which has been open-sourced, the creators are able to offer impressive speed boosts in many cases. The data structures behind the performance gains are also impressive. Acunu also offers some of the traditional benefits of a commercially supported distribution, such as automatic configuration and other administration tools.

Activities

A logical grouping or classification of one or more jobs running on a cluster.

Aggregation

A process of searching, gathering and presenting data

Algorithm

A mathematical formula placed in software that performs an analysis on a set of data.

Anonymization

The severing of links between people in a database and their records to prevent the discovery of the source of the records.

Apache Hadoop

An open source platform that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The platform particularly suited to large volumes of unstructured data such as Facebook comments and Twitter tweets, email and instant messages, and security and application logs.

Apache Spark

An open-source data analytics cluster computing framework, originally developed in the AMPLab at UC Berkeley. It is built on top of the Hadoop Distributed File System and has much faster performance compared to MapReduce. It provides high-level APIs in Scala, Python and Java.

Artificial Intelligence

Developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

Automatic identification and capture (AIDC)

Any method of automatically identifying and collecting data on items, and then storing the data in a computer system. For example, a scanner might collect data about a product being shipped via an RFID chip.

Avro

Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing remote procedure calls. (or)

An Apache™ open source project that provides data serialisation and data exchange services for Hadoop®

Azkaban

The trickiest part of building a working system using these new data tools is the integration. The individual services need to be tied together into sequences of operations that are triggered by your business logic, and building that plumbing is surprisingly time consuming. Azkaban is an open source project from LinkedIn that lets you define what you want to happen as a job flow, possibly made up of many dependent steps, and then handles a lot of the messy housekeeping details. It keeps track of the log outputs, spots errors and emails about errors as they happen, and provides a friendly web interface so you can see how your jobs are getting on. Jobs are created as text files, using a very minimal set of commands, with any complexity expected to reside in the Unix commands or Java programs that the step calls.

Balancer

A service ensuring that all nodes in the cluster store about the same amount of data, within a set range. Data are balanced over the nodes in the cluster, not over the disks in a node.

Behavioral analytics

Using data about people’s behavior to understand intent and predict future actions.

Big Data Scientist

Someone who is able to develop the algorithms to make sense out of big data.

Business Intelligence (BI)

The general term used for the identification, extraction, and analysis of data.

Cascading

Cascading provides a higher level of abstraction for Hadoop, allowing developers to create complex jobs quickly, easily, and in several different languages that run in the JVM, including Ruby, Scala, and more. In effect, this has shattered the skills barrier, enabling Twitter to use Hadoop more broadly.

Call Detail Record (CDR) analysis

CDRs contain data that a telecommunications company collects about phone calls, such as time and length of call. This data can be used in any number of analytical applications.

CAP Theorem

The CAP theorem states that it is impossible to guarantee all 3 features simultaneously in a distributed system:

* Consistency – all nodes see the same data at the same time
* Availability – every request receives a response about whether it succeeded or failed
* Partition tolerance – when parts of the nodes fail, the system continues to operate properly

Now, once you have Big Data, you need a distributed system. At that point, you need to stop thinking of ACID and start deciding which of the 3 properties you can live without (or live with less of).

Cascading

Most real-world Hadoop applications are built of a series of processing steps, and Cascading lets you define that sort of complex workflow as a program. You lay out the logical flow of the data pipeline you need, rather than building it explicitly out of MapReduce steps feeding into one another. To use it, you call a Java API, connecting objects that represent the operations you want to perform into a graph. The system takes that definition, does some checking and planning, and executes it on your Hadoop cluster. There are a lot of built-in objects for common operations like sorting, grouping, and joining, and you can write your own objects to run custom processing code.

Cascalog

Cascalog is a functional data processing interface written in Clojure. Influenced by the old Datalog language and built on top of the Cascading framework, it lets you write your processing code at a high level of abstraction while the system takes care of assembling it into a Hadoop job. It makes it easy to switch between local execution on small amounts of data to test your code and production jobs on your real Hadoop cluster. Cascalog inherits the same approach of input and output taps and processing operations from Cascading, and the functional paradigm seems like a natural way of specifying data flows. It’s a distant descendant of the original Clojure wrapper for Cascading, cascading-clojure.

Cassandra

Cassandra is a distributed and Open Source database. Designed to handle large amounts of distributed data across commodity servers while providing a highly available service. It is a NoSQL solution that was initially developed by Facebook. It is structured in the form of key-value. (or)

An open source database in the form of a key-value store. Fast on read and write, good high availability. Crappy when it comes to non-simple schemas and the ability to do anything resembling SQL stuff. Consistency also isn’t one of its strong points (at least not for something like banking applications).

CDH

Cloudera's Distribution including Apache Hadoop, the version of Apache Hadoop and related components installed on Oracle Big Data Appliance.

Cell phone data

Cell phones generate a tremendous amount of data, and much of it is available for use with analytical applications.

Chukwa

Chukwa is a Hadoop subproject devoted to large-scale log collection and analysis. Chukwa is built on top of the Hadoop distributed filesystem (HDFS) and MapReduce framework and inherits Hadoop’s scalability and robustness. Chukwa also includes a flexible and powerful toolkit for displaying monitoring and analyzing results, in order to make the best use of this collected data.

Classification analysis

A systematic process for obtaining important and relevant information about data, also meta data called; data about data.

Clickstream Analytics

The analysis of users’ Web activity through the items they click on a page.

Client machine

Neither a NameNode or a DataNode, Client machines have Hadoop installed on them. They’re responsible for loading data into the cluster, submitting MapReduce jobs and viewing the results of the job once complete.

Clojure

Clojure is a dynamic programming language based on LISP that uses the Java Virtual Machine (JVM). It is well suited for parallel data processing.

Cloud

A broad term that refers to any Internet-based application or service that is hosted remotely.

Cloud computing

A distributed computing system over a network used for storing data off-premises

Cluster (Hadoop)

A set of nodes configured to work together based on a common Hadoop component stack, with HDFS and MapReduce as the foundation. (or)

A group of servers on a network that are configured to work together. A server is either a master node or a worker node.

Clustering analysis

The process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

Cold data storage

Storing old data that is hardly used on low-power servers. Retrieving the data will take longer

Columnar database or column-oriented database

A database that stores data by column rather than by row. In a row-based database, a row might contain a name, address, and phone number. In a column-oriented database, all names are in one column, addresses in another, and so on. A key advantage of a columnar database is faster hard disk access.

Comparative analysis

It ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

Comparators

Two ways you may compare your keys is by implementing the interface or by implementing the RawComparator interface. In the former approach, you will compare (deserialized) objects, but in the latter approach, you will compare the keys using their corresponding raw bytes.

Components (Hadoop)

The individual installed software products composing a complete Hadoop cluster. Some components are active and include servers, such as HDFS, and some are passive libraries. The servers of active components provide a service.

Components consist of roles that represent the different configurations required by the component. Components have a role on each host server. For example, HDFS roles include namenode, secondary namenode, and datanode.

Components and roles run as named instances in an active cluster, such as NameNode1, DataNode73, and HDFS001.

Complex event processing (CEP)

CEP is the process of monitoring and analyzing all events across an organization’s systems and acting on them when necessary in real time.

Confabulation

The act of making an intuition-based decision appear to be data-based.

Cross-channel analytics

Analysis that can attribute sales, show average order value, or the lifetime value.

Data access

The act or method of viewing or retrieving stored data.

Dashboard

A graphical representation of the analyses performed by the algorithms

Data aggregation

The act of collecting data from multiple sources for the purpose of reporting or analysis.

Data architecture and design

How enterprise data is structured. The actual structure or design varies depending on the eventual end result required. Data architecture has three stages or processes: conceptual representation of business entities. the logical representation of the relationships among those entities, and the physical construction of the system to support the functionality.

Database

A digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system (DBMS).

Database administrator (DBA)

A person, often certified, who is responsible for supporting and maintaining the integrity of the structure and content of a database.

Database as a service (DaaS)

A database hosted in the cloud and sold on a metered basis. Examples include Heroku Postgres and Amazon Relational Database Service.

Database management system (DBMS)

Software that collects and provides access to data in a structured format.

Data center

A physical facility that houses a large number of servers and data storage devices. Data centers might belong to a single organization or sell their services to many organizations.

Data cleansing

The act of reviewing and revising data to remove duplicate entries, correct misspellings, add missing data, and provide more consistency.

Data collection

Any process that captures any type of data.

Data custodian

A person responsible for the database structure and the technical environment, including the storage of data.

Data-directed decision making

Using data to support making crucial decisions.

Data exhaust

The data that a person creates as a byproduct of a common activity–for example, a cell call log or web search history.

Data feed

A means for a person to receive a stream of data. Examples of data feed mechanisms include RSS or Twitter.

Data governance

A set of processes or rules that ensure the integrity of the data and that data management best practices are met.

Data integration

The process of combining data from different sources and presenting it in a single view.

Data integrity

The measure of trust an organization has in the accuracy, completeness, timeliness, and validity of the data.

Data mart

The access layer of a data warehouse used to provide data to users.

Data migration

The process of moving data between different storage types or formats, or between different computer systems.

Data mining

The process of deriving patterns or knowledge from large data sets.

Data model, data modeling

A data model defines the structure of the data for the purpose of communicating between functional and technical people to show data needed for business processes, or for communicating a plan to develop how data is stored and accessed among application development team members.

DataNode

Runs on “slave nodes,” which make up the majority of the machines within a cluster. The NameNode instructs data files to be split into blocks, each of which are replicated three times and stored on machines across the cluster. These replicas ensure the entire system won’t go down if one server fails or is taken offline—known as “fault tolerance.”

Data point

An individual item on a graph or a chart.

Data profiling

The process of collecting statistics and information about data in an existing source.

Data quality

The measure of data to determine its worthiness for decision making, planning, or operations.

Data replication

The process of sharing information to ensure consistency between redundant sources.

Data repository

The location of permanently stored data.

Data science

A recent term that has multiple definitions, but generally accepted as a discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning, and database engineering to solve complex problems.

Data scientist

A practitioner of data science.

Data security

The practice of protecting data from destruction or unauthorized access.

Data set

A collection of data, typically in tabular form.

Data source

Any provider of data–for example, a database or a data stream.

Data steward

A person responsible for data stored in a data field.

Data structure

A specific way of storing and organizing data.

Data visualization

A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

Data warehouse

A place to store data for the purpose of reporting and analysis.

De-identification

The act of removing all data that links a person to a particular piece of information.

Demographic data

Data relating to the characteristics of a human population.

Deep Thunder

IBM’s weather prediction service that provides weather data to organizations such as utilities, which use the data to optimize energy distribution.

Distributed cache

A data cache that is spread across multiple systems but works as one. It is used to improve performance.

Distributed object

A software module designed to work with other distributed objects stored on other computers.

Distributed processing

The execution of a process across multiple computers connected by a computer network.

Distributed File System

Systems that offer simplified, highly available access to storing, analysing and processing data

Document Store Databases

A document-oriented database that is especially designed to store, manage and retrieve documents, also known as semi structured data.

Document management

The practice of tracking and storing electronic documents and scanned images of paper documents.

Drill

An open source distributed system for performing interactive analysis on large-scale datasets. It is similar to Google’s Dremel, and is managed by Apache.

Driver Jar File

Jar file that will allow Hadoop to drive an H2O launch as well as create a connection between HDFS and H2O for importing data from HDFS.

Elasticsearch

An open source search engine built on Apache Lucene.

Event analytics

Shows the series of steps that led to an action.

Exabyte

One million terabytes, or 1 billion gigabytes of information.

External data

Data that exists outside of a system.

Extract, transform, and load (ETL)

A process used in data warehousing to prepare data for use in reporting or analytics.

Exploratory analysis

Finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.

Failover

The automatic switching to another computer or node should one fail.

Flume

A service for collecting, aggregating, and moving large amounts of log and event data into Hadoop. (or)

Flume is a framework for populating Hadoop with data. Agents are populated throughout ones IT infrastructure – inside web servers, application servers and mobile devices, for example – to collect data and integrate it into Hadoop.

Greenplum

Though not strictly a NoSQL database, the Greenplum system offers an interesting way of combining a flexible query language with distributed performance. Built on top of the Postgres open source database, it adds in a distributed architecture to run on a cluster of multiple machines, while retaining the standard SQL interface. It automatically shards rows across machines, by default based on a hash of a table’s primary key, and works to avoid data loss both by using RAID drive setups on individual servers and by replicating data across machines. It’s normally deployed on clusters of machines with comparatively fast processors and large amounts of RAM, in contrast to the pattern of using commodity hardware that’s more common in the web world.

Grid computing

The performing of computing functions using resources from multiple distributed systems. Grid computing typically involves large files and are most often used for multiple applications. The systems that comprise a grid computing network do not have to be similar in design or in the same geographic location.

Graph Databases

They use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbour element.

H2O

H2O makes Hadoop do math. H2O is an Apache v2 licensed open source math and prediction engine.

H2O Cluster

A group of H2O nodes that operate together to work on jobs. H2O scales by distributing work over many H2O nodes. (Note multiple H2O nodes can run on a single Hadoop node if sufficient resources are available.) All H2O nodes in an H2O cluster are peers. There is no “master” node.

H2O Key Value

H2O implements a distributed in-memory Key/Value store within the H2O cluster. H2O uses Keys to uniquely identify data sets that have been read in (pre-parse), data sets that have been parsed (into HEX format), and models (e.g. GLM) that have been created. For example, when you ingest your data from HDFS into H2O, that entire data set is referred to by a single Key.

H2O Node

H2O nodes are launched via Hadoop MapReduce and run on Hadoop DataNodes. (At a system level, an H2O node is a Java invocation of h2o.jar.) Note that while Hadoop operations are centralized around HDFS file accesses, H2O operations are memory-based when possible for best performance. (H2O reads the dataset from HDFS into memory and then attempts to perform all operations to the data in memory.)

Hadoop

An open source software library project administered by the Apache Software Foundation. Apache defines Hadoop as “a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model.” (or)

An open source big-data platform. Cloudera, MapR, and Hortonworks are distro providers of Hadoop. Data is stored in HDFS (DataNode, NameNode) and processed through MapReduce and managed via JobTracker.

Hadoop Common

Usually only referred to by programmers, Hadoop Common is a common utilities library that contains code to support some of the other modules within the Hadoop ecosystem. When Hive and HBase want to access HDFS, for example, they do so using JARs (Java archives), which are libraries of Java code stored in Hadoop Common.

Hama

Hama is a distributed computing framework based on Bulk Synchronous Parallel computing techniques for massive scientific computations e.g., matrix, graph and network algorithms. It’s a Top Level Project under the Apache Software Foundation.

HANA

A software/hardware in-memory computing platform from SAP designed for high-volume transactions and real-time analytics.

HBase

HBase is a non-relational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily. (or)

An open-source, distributed, versioned, non-relational database modeled after Google’s Bigtable: A Distributed Storage System for Structured Data. (or)

HBase is a columnar database management system that is built on top of Hadoop and runs on HDFS. Like MapReduce, HBase applications are written in Java, as well as other languages via their Thrift database, which is a framework that allows cross-language services development. The key difference between MapReduce and HBase is that HBase is intended to work with random workloads.

HCatalog

HCatalog is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored. (or)

Table and storage management service for Hadoop data that presents a table abstraction so that you do not need to know where or how your data is stored.

HDFS (Hadoop Distributed File System)

HDFS (Hadoop Distributed File System) the storage layer of Hadoop, is a distributed, scalable, Java-based file system adept at storing large volumes of unstructured. (or)

An acronym for “Hadoop Distributed File System”, which breaks large application workloads into smaller data blocks that are replicated and distributed across a cluster of commodity hardware for faster processing. (or)

Hadoop Distributed File System is a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.

HEX format

The HEX format is an efficient internal representation for data that can be used by H2O algorithms. A data set must be parsed into HEX format before you can operate on it.

Hive

A data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. It allows you to query data using a SQL-like language called HiveQL (HQL). (or)

Hive is a Hadoop-based data warehousing-like framework originally developed by Facebook. It allows users to write queries in a SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, etc. (or)

Hive allows users who aren’t familiar with programming to access and analyze big data in a less technical way, using a SQL-like syntax called Hive Query Language (HiveQL). HiveQL is used to create programs that run just like MapReduce would on a cluster.

In a very general sense, Hive is used for complex, long-running tasks and analyses on large sets of data, e.g. analyzing the performance of every store within a particular region for a chain retailer.

HiveQL (HQL)

A SQL like query language for Hadoop used to execute MapReduce jobs on HDFS.

Hive Thrift

A remote procedure call (RPC) interface for remote access to CDH for Hive queries.

HotSpot

A Java Virtual Machine (JVM) that is maintained and distributed by Oracle. It automatically optimizes code that executes frequently, leading to high performance. HotSpot is the standard JVM for the other components of the Oracle Big Data Appliance stack.

Hosts

Devices, such as a computer or a switch, attached to a computer or telecommunications network, or a point in a network topology where lines intersect or branch.

Hue

A browser-based desktop interface for interacting with Hadoop. (or)

Hue (Hadoop User Experience) is an open source web-based interface for making it easier to use Apache Hadoop. It features a file browser for HDFS, an Oozie Application for creating workflows and coordinators, a job designer/browser for MapReduce, a Hive and Impala UI, a Shell, a collection of Hadoop API and more.

Impala

An SQL query engine with massive parallel processing (MPP) power, running natively on the Apache Hadoop framework. It shares the same flexible file system (HDFS), metadata, resource management and security frameworks as used by other Hadoop ecosystem components. (or)

Impala (By Cloudera) provides fast, interactive SQL queries directly on your Apache Hadoop data stored in HDFS or HBase using the same metadata, SQL syntax (Hive SQL), ODBC driver and user interface (Hue Beeswax) as Apache Hive. This provides a familiar and unified platform for batch-oriented or real-time queries.

In-database analytics

The integration of data analytics into the data warehouse.

In-Memory

There’s a whole set of databases that don’t use any storage – or at least not for their daily operations. These are in-memory databases. They are fast as hell, scale horizontally up to a point most of the time, and are expensive when it comes to hardware (memory costs more than disk).

For real-time stuff, they probably need to be part of the solution.

In-memory database

Any database system that relies on memory for data storage.

In-memory data grid (IMDG)

The storage of data in memory across multiple servers for the purpose of greater scalability and faster access or analytics.

Internet of Things

Ordinary devices that are connected to the internet at any time any where via sensors

JAQL

Query language designed for JavaScript Object Notation (JSON). Primarily used to analyse large-scale semi-structured data. Core features include user extensibility and parallelism.

Job

A mapper or reducer execution across a dataset. A job may split data to be processed across mapper tasks for parallel processing, with a master (JobTracker) scheduling and monitoring jobs across slaves (TaskTracker).

JobTracker

The service within Hadoop which distributes MapReduce tasks to specific nodes in the cluster. (or)

The JobTracker oversees how MapReduce jobs are split up into tasks and divided among nodes within the cluster. (or)

The JobTracker is the service within Hadoop that farms out MapReduce tasks to specific nodes in the cluster.

JobTracker port

Port where you can access the JobTracker. The default port might be different for each distribution.

Kafka

Kafka (developed by LinkedIn) is a distributed publish-subscribe messaging system that offers a solution capable of handling all data flow activity and processing these data on a consumer website. This type of data (page views, searches, and other user actions) are a key ingredient in the current social web.

Kerberos

Network authentication protocol which works on the basis of 'tickets' to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner

Key Value Stores

Key value stores allow the application to store its data in a schema-less way. The data could be stored in a datatype of a programming language or an object. Because of this, there is no need for a fixed data model.

KeyValue Databases

They store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

Latency

Any delay in a response or delivery of data from one point to another.

Linked data

As described by World Wide Web inventor Time Berners-Lee, “Cherry-picking common attributes or languages to identify connections or relationships between disparate sources of data.”

Load balancing

The process of distributing workload across a computer network or computer cluster to optimize performance.

Location analytics

Location analytics brings mapping and map-driven analytics to enterprise business systems and data warehouses. It allows you to associate geospatial information with datasets.

Location data

Data that describes a geographic location.

Log file

A file that a computer, network, or application creates automatically to record events that occur during operation–for example, the time a file is accessed.

Machine-generated data

Any data that is automatically created from a computer process, application, or other non-human source.

Machine2Machine data

Two or more machines that are communicating with each other

Machine learning

The use of algorithms to allow a computer to analyze data for the purpose of “learning” what action to take when a specific pattern or event occurs.

Mapper Size

The memory allocated to each mapper task that will launch on each of the Hadoop Nodes.

MapR

MapR is a commercial distribution of Hadoop aimed at enterprises. It includes its own file systems that are a replacement for HDFS, along with other tweaks to the framework, like distributed name nodes for improved reliability. The new file system aims to offer increased performance, as well as easier backups and compatibility with NFS to make it simpler to transfer data in and out. The programming model is still the standard Hadoop one; the focus is on improving the infrastructure surrounding the core framework to make it more appealing to corporate customers.

MapReduce

A software framework for easily writing applications that process vast amounts of data (multi-terabyte data-sets) in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. Hadoop acts as a platform for executing MapReduce. (or)

MapReduce is a software framework that serves as the compute layer of Hadoop. MapReduce jobs are divided into two (obviously named) parts. The “Map” function divides a query into multiple parts and processes data at the node level. The “Reduce” function aggregates the results of the “Map” function to determine the “answer” to the query.

Mashup

The process of combining different datasets within a single application to enhance output–for example, combining demographic data with real estate listings.

Mahout

Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing and statistical modeling and implements them using the Map Reduce model.

Metadata

Data about data; gives information about what the data is about.

MongoDB

A hyped database that wants to replace SQL databases, but probably can’t for a lot of the use cases. It is marketed as being super-fast. If you start digging deeper about it, it seems like an unreliable database to use for mission critical stuff.

(or)

MongoDB is a NoSQL database oriented to documents, developed under the open source concept. It saves data structures in JSON documents with a dynamic scheme (called MongoDB BSON format), making the integration of the data in certain applications more easily and quickly.

MPP database

A database optimized to work in a massively parallel processing environment.

mrjob

Mrjob is a framework that lets you write the code for your data processing, and then transparently run it either locally, on Elastic MapReduce, or on your own Hadoop cluster. Written in Python, it doesn’t offer the same level of abstraction or built-in operations as the Java-based Cascading. The job specifications are defined as a series of map and reduce steps, each implemented as a Python function. It is great as a framework for executing jobs, even allowing you to attach a debugger to local runs to really understand what’s happening in your code.

Multi-Dimensional Databases

A database optimized for data online analytical processing (OLAP) applications and for data warehousing.

MultiValue Databases

They are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

MySQL Server

A SQL-based relational database management system. Cloudera Manager, Oracle Data Integrator, Hive, and Oozie use MySQL Server as a metadata repository on Oracle Big Data Appliance.

NameNode

The core of the HDFS file system. The NameNode maintains a record of all files stored on the Hadoop cluster. (or)

A service that maintains a directory of all files in HDFS and tracks where data is stored in the cluster. Maintaining master-to-slave data nodes.

Network analysis

Viewing relationships among the nodes in terms of the network or graph theory, meaning analysing connections between nodes in a network and the strength of the ties.

NewSQL

An elegant, well-defined database system that is easier to learn and better than SQL. It is even newer than NoSQL. (or)

NewSQL is a marketing term that started due to NoSQL. The old regime of SQL databases had to reinvent itself, so NewSQL came to being. Other than that, I have no real clue what to do with this NewSQL thing.

Nodes

An abstract unit that composes a cluster; a vertex in a graph.

NoSQL

NoSQL (commonly interpreted as “not only SQL“) is a broad class of database management systems identified by non-adherence to the widely used relational database management system model. NoSQL databases are not built primarily on tables, and generally do not use SQL for data manipulation.

OASM

Oracle Automated Service Manager, a service for monitoring the health of Oracle Sun hardware systems. Formerly named Sun Automatic Service Manager (SASM).

Object Databases

They store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

Object-based Image Analysis

Analysing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.

Online analytical processing (OLAP)

The process of analyzing multidimensional data using three operations: consolidation (the aggregation of available), drill-down (the ability for users to see the underlying details), and slice and dice (the ability for users to select subsets and view them from different perspectives).

Online transactional processing (OLTP)

The process of providing users with access to large amounts of transactional data in a way that they can derive meaning from it.

Oozie

A workflow engine for Hadoop. (or)

Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages – such as Map Reduce, Pig and Hive — then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

OpenDremel

The open source version of Google’s Big Query java code. It is being integrated with Apache Drill.

Open Data Center Alliance (ODCA)

A consortium of global IT organizations whose goal is to speed the migration of cloud computing.

Operational data store (ODS)

A location to gather and store data from multiple sources so that more operations can be performed on it before sending to the data warehouse for reporting.

Parallel data analysis

Breaking up an analytical problem into smaller components and running algorithms on each of those components at the same time. Parallel data analysis can occur within the same system or across multiple systems.

Parallel method invocation (PMI)

Allows programming code to call multiple functions in parallel.

Parallel processing

The ability to execute multiple tasks at the same time.

Parallel query

A query that is executed over multiple system threads for faster performance.

Parse

The parse operation converts an in-memory raw data set (in CSV format, for example) into a HEX format data set. The parse operation takes a data set named by a Key as input, and produces a HEX format Key,Value output.

Pattern recognition

The classification or labeling of an identified pattern in the machine learning process.

Pentaho

Pentaho offers a suite of open source Business Intelligence (BI) products called Pentaho Business Analytics providing data integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities

Petabyte

One million gigabytes or 1,024 terabytes.

Pig

A high level programming language for creating MapReduce programs used within Hadoop. (or)

Pig Latin is a Hadoop-based language developed by Yahoo. It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

Predictive analytics

Using statistical functions on one or more datasets to predict trends or future events.

Predictive modeling

The process of developing a model that will most likely predict a trend or outcome.

Protocol Buffers

An open sourced version of the system that Google uses internally on most of its projects, the Protocol Buffers stack is an IDL similar to Thrift. One difference is that Thrift includes network client and server code in its generated stubs, whereas protobuf limits its scope to pure serialization and deserialization. The biggest differentiator between the two projects is probably their developer base. Though the code is open source, Google is the main contributor and driver for Protocol Buffers, whereas Thrift is more of a classic crowd-sourced project. If your requirements skew towards stability and strong documentation, Protocol Buffers is going to be attractive, whereas if you need a more open, community-based approach, Thrift will be a lot more appealing.

Public data

Public information or data sets that were created with public funding

Puppet

A configuration management tool for deploying and configuring software components across a cluster. The Oracle Big Data Appliance initial software installation uses Puppet.

The Puppet tool consists of these components: puppet agents, typically just called puppets; the puppet master server; a console; and a cloud provisioner

puppet agent

A service that primarily pulls configurations from the puppet master and applies them. Puppet agents run on every server in Oracle Big Data Appliance.

puppet master

A service that primarily serves configurations to the puppet agents.

Query

Asking for information to answer a certain question

Query analysis

The process of analyzing a search query for the purpose of optimizing it for the best possible result.

R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

Re-identification

Combining several data sets to find a certain person within anonymized data

Real-time data

Data that is created, processed, stored, analysed and visualized within milliseconds

Recommendation engine

An algorithm that analyzes a customer’s purchases and actions on an e-commerce site and then uses that data to recommend complementary products.

Reference data

Data that describes an object and its properties. The object may be physical or virtual.

Risk analysis

The application of statistical methods on one or more datasets to determine the likely risk of a project, action, or decision.

Root-cause analysis

The process of determining the main cause of an event or problem.

Routing analysis

Finding the optimized routing using many different variables for a certain means of transport in order to decrease fuel costs and increase efficiency.

S4

Yahoo! initially created the S4 system to make decisions about choosing and positioning ads, but the company open sourced it after finding it useful for processing arbitrary streams of events. S4 lets you write code to handle unbounded streams of events, and runs it distributed across a cluster of machines, using the ZooKeeper framework to handle the housekeeping details. You write data sources and handlers in Java, and S4 handles broadcasting the data as events across the system, load-balancing the work across the available machines. It’s focused on returning results fast, with low latency, for applications like building near real-time search engines on rapidly changing content. This sets it apart from Hadoop and the general MapReduce approach, which involves synchronization steps within the pipeline, and so some degree of latency. One thing to be aware of is that S4 uses UDP and generally offers no delivery guarantees for the data that’s passing through the pipeline. It usually seems possible to adjust queue sizes to avoid data loss, but it does put the burden of tuning to reach the required level of reliability on the application developer.

Scalability

The ability of a system or process to maintain acceptable performance levels as workload or scope increases.

Schema

The structure that defines the organization of data in a database system.

Search data

Aggregated data about search terms used over time.

Semi-structured data

Data that is not structured by a formal data model, but provides other means of describing the data and hierarchies.

Sentiment analysis

The application of statistical functions on comments people make on the web and through social networks to determine how they feel about a product or company.

Server

A physical or virtual computer that serves requests for a software application and delivers those requests over a network.

Slot

A map or reduce computation unit on a node. Each active map or reduce task occupies one slot, which could be a map or a reduce slot. A TaskTracker has a configured number of slots available (mapred.tasktracker.{map,reduce}.tasks.maximum in mapred-site.xml), and JobTracker allocates work to the TaskTracker with available slots nearest to the data.

Cluster metrics include the cluster slot capacity and the number of currently allocated map and reduce slots.

Solr/Lucene

Enterprise search tool from the Apache Lucene project that offers powerful search tools, including hit highlighting, as well as indexing capabilities, reliability and scalability, a central configuration system, and failover and recovery.

Spatial analysis

It refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.

Spilling

An H2O node may choose to temporarily “spill” data from memory onto disk. (Think of this like swapping.) In Hadoop environments, H2O spills to HDFS. Usage is intended to function like a temporary cache, and the spilled data is discarded when the job is done.

SQL

A programming language for retrieving data from a relational database

Sqoop

A tool designed to transfer data between Hadoop and relational databases. (or)

ELT tool to support transfer of data between Hadoop and structured data sources. (or)

Sqoop is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop. It allows users to specify the target location inside of Hadoop and instruct Sqoop to move data from Oracle, Teradata or other relational databases to the target.

Stack (Hadoop)

Hadoop software layers with which the Splunk App for HadoopOps interacts:

* Data processing layer; encapsulates the MapReduce framework.
* Data storage layer; encapsulates the file system (HDFS).

Storm

Storm is a system of real-time distributed computing, open source and free, born into Twitter. Storm makes it easy to reliably process unstructured data flows in the field of real-time processing, which made Hadoop for batch processing.

Software as a service (SaaS)

Application software that is used over the web by a thin client or web browser. Salesforce is a well-known example of SaaS.

Storage

Any means of storing data persistently.

Storm

An open-source distributed computation system designed for processing multiple data streams in real time.

Structured data

Data that is organized by a predetermined structure.

Structured Query Language (SQL)

A programming language designed specifically to manage and retrieve data from a relational database system.

Task

A mapper or reducer instance operating on a slice of data. Tasks are executed by the Hadoop TaskTracker, which assigns tasks to nodes with resources available for executing the task. Each active map or reduce task occupies one slot.

TaskAttempt

An instance of a map or reduce task, which is identified by a task ID. The JobTracker may run a task on more than one node, if it either fails or to enable getting faster results from another node; this adds to the number of attempts.

TaskTracker

The TaskTracker accepts tasks from the JobTracker, performs the work and alerts the JobTracker once it’s done. TaskTrackers and DataNodes are located on the same nodes to improve performance

TaskWaiting

A task state of waiting to be launched.

Text analytics

The application of statistical, linguistic, and machine learning techniques on text-based sources to derive meaning or insight.

Thrift

“Thrift is a software framework for scalable cross-language services development. It combines a software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml.”

Transactional data

Data that changes unpredictably. Examples include accounts payable and receivable data, or data about product shipments.

Unstructured data

Data that has no identifiable structure–for example, the text of email messages.

Value

All that available data will create a lot of value for organizations, societies and consumers. Big data means big business and every industry will reap the benefits from big data.

Volume

The amount of data, ranging from megabytes to brontobytes

Visualization

A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

WebHDFS Apache Hadoop

WebHDFS Apache Hadoop provides native libraries for accessing HDFS. However, users prefer to use HDFS remotely over the heavy client side native libraries. For example, some applications need to load data in and out of the cluster, or to externally interact with the HDFS data. WebHDFS addresses these issues by providing a fully functional HTTP REST API to access HDFS.

Weather data

Real-time weather data is now widely available for organizations to use in a variety of ways. For example, a logistics company can monitor local weather conditions to optimize the transport of goods. A utility company can adjust energy distribution in real time.

Whirr

A set of libraries for running cloud services. It’s ideal for running temporary Hadoop clusters to carry out a proof of concept, or to run a few one-time jobs

XML Databases

XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported and serialized into any format needed.

YARN (Yet Another Resource Negotiator)

A resource manager for Hadoop 2. YARN is short for “Yet another resource negotiator”. (or)

A resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users’ applications. (or)

YARN is an updated way of handling the delegation of resources for MapReduce jobs. It takes the place of the JobTracker and TaskTracker. In our house example, if JobTracker and TaskTracker can be thought of as the foreman, YARN is a foreman with an MBA—it’s a more advanced way of carrying out MapReduce jobs. It also gives you added abilities, such as the ability to work with frameworks other than MapReduce and to translate jobs developed in languages other than Java.

ZooKeeper

Allows Hadoop administrators to track and coordinate distributed applications. (or)

ZooKeeper is a software project of the Apache Software Foundation, a service that provides centralized configuration and open code name registration for large distributed systems. ZooKeeper is a subproject of Hadoop.