MAPR ACADEMY ESSENTIALS

ESS 1000 – Big Data Essentials Glossary

Spring 2017

For use with the following courses:

ESS 1000 – Big Data Essentials
ESS 100 – Introduction to Big Data
ESS 101 – Apache Hadoop Essentials
ESS 200 – MapR Administration Essentials
ESS 300 – MapReduce Essentials
ESS 320 – MapR-DB Essentials
ESS 350 – MapR Streams Essentials
ESS 360 – Apache Spark Essentials
ESS 400 – Apache Drill Essentials
ESS 440 – Apache Hive Essentials
ESS 450 – Apache Pig Essentials

Glossary of Big Data Terms

For more definitions related specifically to the MapR Converged Data Platform, visit:
http://maprdocs.mapr.com/51/index.html - ReferenceGuide/Glossary.html

A

ACID – Atomicity, Consistency, Isolation, and Durability. These features are important to ensure data is reliably stored and processed on a database.

ad hoc – A quick query, report, or analysis performed once, typically done by a data analyst as part of the early steps of data exploration.

administrator – A person responsible for managing a cluster. This includes preparing nodes, adding and removing users, configuring security, testing performance against benchmarks, upgrading software, disaster recovery planning, and day-to-day problem solving.

aggregate – A summary of a large amount of data. The mean average of a data set is an example of an aggregate function.

analyst – A person responsible for analyzing data. This includes data mining, extraction, normalization, filtering, aggregation, querying, interpreting, graphing, and making predictions.

ANSI-SQL – A standardized form of SQL, established by the American National Standards Institute. See also: *SQL*.

Apache – A group of open-source software projects, including Drill, Flume, Hadoop, HBase, Hive, Kafka, Pig, Spark, Sqoop, Storm, ZooKeeper, and many others.

API – Application Programming Interface. APIs are used to define how one program or application connects or interacts with another program or application.

architect – A person responsible for managing how data is ingested and stored.

AWS – Amazon Web Services. A collection of virtual machines, running on the cloud, managed by Amazon. AWS is one option for completing the lab exercises offered in many MapR Academy courses.

B

big data – Data characterized by one or more of the following: high volume, high velocity, or high variety. See also: *volume*, *velocity*, *variety*.

Bigtable – A compressed, high performance, and proprietary data storage system built on Google File System. See also: *HBase*, *MapR-DB*.

bit – The smallest unit of data, typically characterized as a binary 0 or 1.

block – A unit of data used by a file system.

byte – A unit of data made up of 8 bits of information. A single typed character, such as "x" or "4" contains about one byte of information.

C

cache – Temporary memory, used while actively processing data.

chunk – The shard size of files in MapR-FS. By default, a chunk is 256 megabytes.

CLDB – Container Location Database. A service that tracks the location of every container in MapR-FS. See also: container.

CLI – Command Line Interface. A method of using a program via text commands, typically typed into a terminal or other shell program such as PuTTy. See also: *GUI*.

cloud – An online, remote storage and processing system.

cluster – A collection of nodes. See also: *node*.

column – A set of data values of the same type, organized vertically. See also: *row*.

compression – The process of making a file smaller in size. A zip file is an example of a compressed file.

consumer – An application that reads or processes data in MapR Streams, such as analytics applications, reporting tools, or enterprise dashboards. See also: *producer*.

container – An abstract entity that stores files and directories in MapR-FS. By default, containers are up to 30 gigabytes in size, are striped across storage pools, and are replicated three times.

CRUD – Create, Read, Update, and Delete. These are the basic functions of a computer database.

custom function – A function which is not part of a standard package of functions, typically programmed by the user. See also: *UDF*.

D

data – A set of values or information.

data analyst – See: *analyst*.

database – An organized collection of data.

data pipeline – The process data must go through, from collection (input) to results (output). The data pipeline may involve administrators, analysts, and developers.

data warehouse – A centralized repository storing data from a variety of sources.

data type – A method of categorizing data. Common data types include integers, strings, timestamps, images, sounds, or structured data like CSV or JSON files.

developer – A person responsible for developing programs. This includes designing, developing, deploying, testing, and maintaining code, typically in Java, Python, or Scala.

distributed – Storage or processing which occurs on more than one machine. See also: *local*.

Drill – Part of the Apache Hadoop Ecosystem. A schema-free SQL query engine for Hadoop, NoSQL, and Cloud Storage. See also: *Apache*, *NoSQL*.

E

ecosystem – A set of open source Apache projects that run on Apache Hadoop.

ETL – Extract, Transform, Load. ETL is a common first step in the data pipeline.

exabyte – A unit of data equivalent to about 10^{18} bytes, or 1000 petabytes.

F

flat-wide – A method of organizing data so there are few rows but many columns. See also: *tall-narrow*.

float – A number, typically with a decimal, such as 3.14.

Flume – Part of Apache Hadoop. A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. See also: *Apache*, *log file*.

G

GCP – Google Cloud Platform. A collection of virtual machines, running on the cloud, managed by Google. GCP is one option for completing the lab exercises offered in many MapR Academy courses.

GFS – Google File System. A distributed file system implemented by Google.

gigabyte – A unit of data equivalent to about 10^{9} bytes, or 1000 megabytes. An hour-long music CD contains about one gigabyte of information.

GUI – Graphical User Interface. Allows users to interact with a program without needing to write code. See also: *CLI.*

H

Hadoop – A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

HBase – Part of the Apache Hadoop ecosystem. An open-source, distributed, versioned, non-relational database modeled after Google's Bigtable. See also: *Apache*, *tables*.

HDFS – Hadoop Distributed File System. A distributed file system implemented by Apache.

Hello World – A basic program, designed to teach programmers the basics of a particular language or tool. Often, this program prints the phrase, "Hello, World!"

Hive – Part of the Apache Hadoop ecosystem. A data warehouse infrastructure that provides data summarization and ad hoc querying. See also: *Apache*, *SQL*.

I

IDE – Integrated Development Environment. A text editor which helps developers write, test, and debug code. Examples of IDEs include Eclipse, Notepad++, and XCode.

ingest – The process of importing, loading, or processing data into a database.

integer – A number, typically without a decimal, such as 4.

I/O – Input/Output.

IoT – Internet of Things. A set of networked sensors on objects, such as cars, watches, or smartphones, which collect and send information.

J

Java – A programming language and computing platform.

JavaScript – The programming language of HTML and the Web.

JDBC – Java Database Connectivity. An application programming interface for the programming language Java. See also: *Java, API*.

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

join – The process of combining two or more tables in a relational database.

JSON – JavaScript Object Notation. A type of semi-structured data used in web applications.

K

Kafka – Part of the Apache Hadoop ecosystem. Publish-subscribe messaging rethought as a distributed commit log. See also: *Apache*, *message*.

key-value – A data representation where the key is unique, often a string or timestamp, and the value or values associated with it can be any datatype.

kilobyte – A unit of data equivalent to about 10^3 bytes, or 1000 bytes. A small plain text file contains about one kilobyte of information.

L

latency – The period of time between the input and output of a program.

library – A collection of functions, data, classes, values, or other code used when developing software.

load balancing – The process of distributing network traffic or processing power across nodes to minimize latency.

local – Storage or processing that occurs on a single machine. See also: *distributed*.

log file – A file that records events in an operating system, such as when a user logs on or starts an application.

logical (location or storage) – Refers to the file path where data is physically stored in the file system, as opposed to its physical location. Data that is logically grouped may not necessarily be stored on the same physical server. See also: *physical*.

M

machine learning – A type of artificial intelligence that uses statistical pattern recognition. Common machine learning algorithms include clustering, classification, and collaborative filtering.

Mahout – Part of the Apache Hadoop ecosystem. A scalable machine learning and data mining library. See also: *Apache*, *machine learning*.

map – The first phase of the MapReduce programming model. In the map phase, data is processed into key-value pairs after being split amongst nodes.

MapR Technologies – A company that provides the only Converged Data Platform that integrates the power of Apache Hadoop and Apache Spark with global event streaming, real-time database capabilities, and enterprise storage.

MapR Converged Data Platform – A full Hadoop stack that includes the MapR File System (MapR-FS), the MapR-DB NoSQL database management system, MapR Streams, the MapR Control System (MCS) user interface, and a full family of Hadoop ecosystem projects.

MapR-DB – An enterprise-grade, high-performance, in-Hadoop, NoSQL database management system.

MapReduce – A programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. The three steps of the MapReduce paradigm are: map, shuffle, reduce.

MapR-FS – A random read-write distributed file system that allows applications to concurrently read and write directly to disk. See also: *HDFS*.

MapR Streams – An integrated publish/subscribe message bus in the MapR Converged Data Platform.

megabyte – A unit of data equivalent to about 10^6 bytes, or 1000 kilobytes. A small image or audio file might contain about one megabyte of information.

memory – The amount of data storage available, especially RAM.

Mesos – Part of the Apache Hadoop ecosystem. Abstracts CPU, memory, storage, and other compute resources away from machines (physical or virtual), enabling fault-tolerant and elastic distributed systems to easily be built and run effectively. See also: *Apache, Myriad*.

message – Key-value pairs in MapR Streams, where keys are optional and values contain the data, which can be text, images, video files, or any other type of data.

metadata – Data which describes other data. Metadata might include information like file size, file location, and data type.

mirror – A copy of an object, such as a disk or volume.

mutability – The ability for a file or other piece of data to be overwritten or modified. See also: *read-write*.

Myriad – Part of the Apache Hadoop ecosystem. Enables the co-existence of Apache Hadoop and Apache Mesos on the physical infrastructure. See also: *Apache*, *Mesos*.

N

NameNode – Part of MapR-DB. Tracks metadata, block information, and locations for all files.

NFS – Network File System. A distributed file system protocol allowing users to access files over a network.

node – A collection of disks and computers used for storing and processing data.

NodeManager – Part of Apache Hadoop. A service that works with the ResourceManager to manage the YARN resource containers that run on each node.

NoSQL – Not Only SQL. An extended version of SQL that can query unstructured data not stored in a relational database, in addition to structured data stored in an RDBMS. See also: *SQL*.

null – An empty value. If there are no results for a query, "null" will be returned rather than returning nothing.

O

Oozie – Part of the Apache Hadoop ecosystem. A workflow scheduler system to manage Apache Hadoop jobs. See also: *Apache*.

open source – Software developed with the rights to study, change, and distribute to anyone and for any purpose. Open source software is developed collaboratively in the open source community.

ODBC – Open Database Connectivity.

P

partition – A part of a table, volume, or other object.

petabyte – A unit of data equivalent to about 10^{15} bytes, or 1000 terabytes. A petabyte can contain about 100 uncompressed copies of the English Wikipedia, including the images, text, links, and history.

permissions – Who has access to different parts of your file system. See also: *administrator*.

Pig – Part of the Apache Hadoop ecosystem. A platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. See also: *ETL*, *Apache*.

physical (location or storage) – Refers to the hardware where data is physically stored in the real world, as opposed to its logical location. Data that is physically stored together may be accessed faster. See also: *logical*.

POSIX – Portable Operating System Interface. POSIX often refers to a set of commands and APIs for running a computer from a CLI.

pool – See: *storage pool*.

producer – An application which generates data in MapR Streams. See also: *consumer*.

processing – Operating on data, via a program or application.

Q

query – A method of searching for data. See also: *SQL*.

R

RAID – Redundant Array of Inexpensive Disks.

RDBMS – Relational Database Management System.

read-only – A piece of data or a file system in which users can view the data, but cannot modify it.

read-write – A piece of data or a file system in which users can both view the data as well as modify it.

reduce – The third and last phase of MapReduce. In this phase, each reducer receives one or more partitions, performs application specific logic on the data, and returns the results.

region – A section of a table in HBase. See also: *tablet*.

regular expression – A sequence of characters that defines a pattern. See also: *query*.

relational database – A type of database where data is organized into relations. See also: *database*.

replica – A copy of data.

replication – The process of copying data.

ResourceManager – Part of Apache Hadoop. Responsible for tracking the resources in a cluster, and scheduling applications. See also: *Hadoop*.

row – A set of values, not necessarily of the same data type, organized horizontally. See also: *column*.

S

sandbox – A virtual machine, often used for testing code or programs.

scalability – The ability of a system to grow to handle more processing and storage.

semi–structured – Data which is tagged, but not organized into rows and columns. XML, HTML, and JSON files are examples of semi-structured data. See also: *structured, unstructured*.

shard – The size files are split into when they are distributed across MapReduce jobs or HDFS nodes.

shuffle – The second phase of MapReduce. In the shuffle phase, results from the map phase are merged and transferred to the reducers.

snapshot – A read-only image of a volume at a specific point in time.

Spark – Part of the Apache Hadoop ecosystem. A fast and general engine for large-scale data processing. See also: *Apache*, *machine learning*.

Sqoop – Part of the Apache Hadoop ecosystem. A tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. See also: *Apache*, *ETL*.

Storm – Part of the Apache Hadoop ecosystem. A free and open source distributed real-time computation system. See also: *Apache, ETL*.

storage – The hard drives, nodes, and clusters where data is kept.

storage pool – A group of disks that MapR-FS writes data to.

string – One or more characters of text data, often designated by quotation marks.

SCP – Secure copy. A Unix command used to copy files from one location to another.

SQL – Structured Query Language. A group of related programming languages, including MySQL, SQLite, HiveQL, and ANSI-SQL, which are used to query relational databases.

SSH – Secure shell. A Unix command used to access a virtual machine.

stream (type of data) – A sequence of data, often sent in real-time. See also: *IoT*.

stream (MapR Streams) – A collection of topics in MapR Streams that can be managed together.

structured – Data which is organized into rows and columns. Examples of structured data include CSV files and HBase tables. See also: *semi-structured, unstructured*.

T

tablet – A section of a table in Bigtable or MapR-DB.

tall-narrow – A method of organizing data so there are few columns but many rows. See also: *flat-wide*.

terabyte – A unit of data equivalent to about 10^{12} bytes, or 1000 gigabytes. A terabyte can store about 40 high-definition Bluray discs of information.

throughput – The rate at which data passes through a system or process.

topic – A logical collection of messages in MapR Streams.

topology – The way physical nodes or logical file systems are organized.

U

UDF – User-defined function. UDFs can be created to perform custom operations in many programming environments.

unstructured – Data which is not organized into columns and rows. Audio and video files are common examples of unstructured data. See also: *structured*.

V

variety – The types of data generated. For example, data might be integers, strings, images, sound, video, raw data from sensors, or JSON files from websites.

velocity – The speed at which data is generated. For example, the Internet of Things may generate thousands of data points per second.

virtual machine – An emulation of a computer system, often run through a CLI.

visualization – The process of turning data into graphs, inforgraphics, dashboards, or other presentations. See also: *analyst*.

volume (type of data) – The amount of data generated. For example, the entire English Wikipedia data dump is several gigabytes of just text data.

volume (MapR-FS) – A management entity that stores and organizes containers in MapR-FS; used to distribute metadata, set permissions on data in the cluster, and for data backup.

W

white paper – A concise report, often about a new technology or business issue.

X

XML – Extensible Markup Language. XML is a type of semi-structured data.

Y

YARN – Yet Another Resource Negotiator. Part of Apache Hadoop. Apache YARN is sometimes called MapReduce 2.0. Apache YARN decouples resource management and data processing in Hadoop.

yottabyte – A unit of data equivalent to about 10^{24} bytes, or 1000 zettabytes.

Z

zettabyte – A unit of data equivalent to about 10^{21} bytes, or 1000 exabytes.

ZooKeeper – Part of the Apache Hadoop ecosystem. A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. See also: *Apache*, *administrator*.