Data Science

Big Data	A term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Can be analyzed for insights that lead to better decisions and strategic business moves.
Big Data tools	Hadoop, Hive, Pig, Apache HBase, Cassandra, MapReduce (method), Spark.
Ambari	It is a software project for easier Hadoop management. It enables system administrators to provide, manage and monitor Hadoop clusters. It also provides features for installation and configuration of Hadoop services and a dashboard for monitoring cluster status.
Apache Avro	A data serialization system that relies on schemas for reading data. Using a schema helps cut back on serialization size. Avro also provides data structures, remote procedure call, compact binary data format and integration with dynamic languages.
Apache Flink	A general-purpose data processing platform and a top-level Apache project. It provides efficient, fast, accurate, and fault tolerant handling of massive streams of events. Flink is usable for dozens of big data scenarios and capable of running in standalone mode. Its defining feature is its ability to process streaming data in real time.
Apache Hive	A data warehouse infrastructure built on top of Hadoop. It provides tools to enable easy data ETL, a mechanism to put structures on the data, and the capability for querying and analysis of large datasets stored in Hadoop files.
Apache Mahout	A machine learning framework for creating scalable applications, that can be used by data scientists, mathematicians and statisticians to implement their algorithms. Mahout also offers core algorithms that can be used for classification, clustering and batch based filtering.
Apache Pig	A programming framework used to analyze and transform large data sets. Apache Pig provides a high-level language known as Pig Latin which helps Hadoop developers write data analysis programs. By using various operators provided by the Pig Latin language, programmers can develop their own functions for reading, writing, and processing data.
Apache Spark	An open-source lightning-fast cluster computing technology, designed for fast computation. Has in-memory cluster computing that increases the processing speed of an app.
Apache Flume	A distributed, reliable and high availability service for collecting, accumulating and moving to a centralized repository of large amounts of streaming data from multiple sources.
Apache Zeppelin	A web-based notebook tool for interactive data analytics. Zeppelin supports different technologies that aid in analytics, such as SQL, Python and Apache Spark. Aside from analytics, it can also perform discovery, ingestion, collaboration and visualisation.
Caffe	A deep learning framework, best used in image classification and segmentation, where speed, modularity and expression are important. Caffe can be implemented in different scale projects, from academic to industrial, as it can process more than 60M images in a day.
Hadoop	An open-source software framework that is used for distributed storage and processing of big data sets across clusters of computers using simple programming models; the Apache project.
HBase	A distributed, versioned, non-relational database modeled after Google's Bigtable. It is built on top of HDFS and allows to perform read/write operations on large datasets in real time using Key/Value data. The programming language of HBase is Java. Today HBase is an integral part of the Apache Software Foundation and the Hadoop ecosystem.
HDFS	Hadoop Distributed File System, HDFS for short, is a Java-based distributed file system that allows to store large data sets (files which are in the range of terabytes and petabytes) reliably. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. It is the primary storage used by Hadoop applications.
HDP	Hortonworks Data Platform is a secure, enterprise-ready open source Hadoop distribution based on a centralized architecture (YARN). HDP enables enterprises to deploy, integrate and work with unprecedented volumes of structured and unstructured data.
Impala	A modern, massively distributed SQL query engine for Apache Hadoop. It allows you to analyze, transform and combine data from a variety of data sources. With Impala, you can query data, whether stored in HDFS or HBase, in real time.
Jaql	A functional language designed for processing data and JSON queries on big data. It is suitable for any volume of data, both structured and unstructured. Jaql also works on other data formats, such as XML and CSV, and it is compatible with SQL structured data.
Kylin	A distributed analytics engine that provides a SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets.
MapReduce	

The heart of Apache Hadoop. A software framework for easily writing applications which process vast amounts of data (multi-terabyte datasets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a

MXNet	an acceleration library that helps save time on building and deploying large-scale DNNs. It also offers predefined layers and tools for coding your own, for specifying data structure placement and automating calculations.
Oozie	A workflow scheduler system designed to manage Hadoop jobs. Oozie allows to automates commonly performed tasks. By using it, you can describe workflows to be performed on a Hadoop cluster, schedule those workflows to execute under a specified condition, and even combine multiple workflows and schedules together into a package to manage their full lifecycle.
SnapLogic eXtreme	A big data transformations tool that can process large volumes of information and doesn't require a special set of skills. The tool assists in cost and risk reduction and data analytics. It can be integrated with other Big Data tools and frameworks, including Amazon Elastic MapReduce and Azure HDInsights.
Sqoop	A Java-based tool used for transferring bulk data between Apache Hadoop and structured datastores such as relational databases.