Big Data Technology Overview

Term	Description	See Also
3 Vs	Volume, velocity and variety. And some expand the definition further to include veracity and value as well.	• Big Data - the 5 Vs Everyone Must Know • 5 Vs of Big Data
Agile	*From Wikipedia,* **"**Agile software development is a group of software development methods based on iterative and incremental development, where requirements and solutions evolve through collaboration between self-organizing, cross-functional teams. It promotes adaptive planning, evolutionary development and delivery, a time-boxed iterative approach, and encourages rapid and flexible response to change. It is a conceptual framework that promotes foreseen tight iterations throughout the development cycle."	• The Agile Manifesto
Avro	A data serialization system. *From Wikepedia,* "It is a remote procedure call and serialization framework developed within Apache's Hadoop project. It uses JSON for defining data types and protocols, and serializes data in a compact binary format."	• Apache Avro
Big Insights	BigInsights Enterprise Edition provides a spreadsheet-like data analysis tool to help organizations store, manage, and analyze big data.	• IBM Infosphere Biginsights
Cassandra	A scalable multi-master database with no single points of failure. It provides scalability and high availability without compromising performance.	• Apache Cassandra
Cloudera	Cloudera Inc. is an American-based software company that provides Apache Hadoop-based software, support and services, and training to business customers.	• Cloudera
Data science	The study of the generalizable extraction of knowledge from data	• Wikipedia - Data Science • IBM - Data Scientist • Coursera

Big Data Technology Overview		
Term	Description	See Also
Dremel	Distributed system developed at Google for interactively querying large datasets. It empowers business analysts and makes it easy for business users to access the data rather than having to rely on data engineers. (*Being merged with Apache Drill*)	• Dremel • Google Research
Drill	Apache Drill is the open source version of what Google is doing with Dremel. It empowers business analysts and makes it easy for business users to access the data rather than having to rely on data engineers. It makes large-scale, ad-hoc querying of data possible, making it suitable for data exploration. It makes it possible to scan over petabytes of data in seconds, to answer ad hoc queries and develop compelling visualizations.	• Apache Drill • Community Resources Apache Drill
Eclipse	Eclipse is a popular IDE donated by IBM to the open source community.	• Eclipse
Flume	Open source software developed by Cloudera. It is a facility for collecting and loading data into Hadoop. In Flume, you work sources, decorators, and sinks. A source is any data source. A sink is the target of a specific operation. A decorator is an operation on the stream that can transform the stream in some manner (i.e., to compress or uncompress data, modify data by adding or removing pieces of information, etc.)	• Apache Flume • Cloudera Flume User Guide
Giraph	Giraph helps empower graph analysis and is often used coupled with graph databases. It is currently used at Facebook to analyze the social graph formed by users and their connections.	• Apache Giraph
Graph databases	Graph databases are used when you want to take a graph approach rather than a relational approach. They store data in a graph and are capable of elegantly representing any kind of data in a highly accessible way.	• Wikipedia Graph Database

Big Data Technology Overview		
Term	Description	See Also
Gremlin	Often used coupled with graph databases, it is a graph traversal language that helps empower graph analysis.	• Gremlin Wiki GitHub
Hadoop	The core platform for structuring big data.	• Apache Hadoop
HBase	Hbase is the Hadoop database. It is a scalable, distributed database that supports structured data storage for large tables.	• Apache HBase
HDFS	Hadoop Distributed File System	• Apache Hadoop HDFS • IBM Hadoop HDFS
Hive	The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides data warehousing tools to extract, transform and load data, and query this data stored in Hadoop files. It provides data summarization and ad hoc querying.	• Hive Apache
Impala	Cloudera's open source a query engine that runs on Apache Hadoop.	• Cloudera Impala • Wikipedia Cloudera Impala
InfoSphere Streams	IBM product that can be used to quickly determine customer sentiment for a new product based on social media comments. It offers a complete stream computing solution to enable real-time analytic processing of data in motion.	• IBM Infosphere Streams
Iterative methodology	An incremental build approach to develop a system through repeated cycles (iterative) and in smaller portions at a time (incremental), allowing software developers to take advantage of what was learned during development of earlier parts or versions of the system.	• Wikipedia Iterative & Incremental Development

Big Data Technology Overview		
Term	Description	See Also
Jaql	A query language for JavaScript open notation.	• Google - Jaql • IBM - Jaql • Wikipedia - Jaql
Kafka	A high-throughput distributed messaging system. When used with Storm, you can conduct reliable stream processing at a linear scale, assured that every message gets processed in real-time.	• Apache Kafka
Lucene	A text search engine library written in Java.	• Apache Lucene
MapReduce	A YARN-based system for parallel processing of large data sets.	• Google Research Map Reduce • Wikipedia MapReduce • IBM MapReduce
Neo4J	A highly scalable, robust (fully ACID) native graph database.	• Neo4J
NoSQL	A NoSQL or Not Only SQL database does not adhere to the traditional relational database management system (RDMS) structure. It is not built on tables and does not employ SQL to manipulate data. However, it does allow SQL-like query languages to be used.	• Wikipedia NoSQL
Pig	Pig is a platform for analyzing large data sets. It is a high-level data-flow language for parallel computation and expressing data analysis. It was initially developed at Yahoo! To allow Hadoop users to focus more on analyzing large data sets and spend less time having to write mapper and reducer programs.	• Apache Pig • IBM Pig
Oozie	Server-based Workflow scheduler system to manage Apache Hadoop jobs.	• Apache Oozie
Progressive elaboration	Progressive elaboration is a project management term that applies whenever you start a new project and are unsure what the outcome will be. Over time, you gather information and adapt your strategy accordingly, continuously improving a plan as more information	• eHow Progressive Elaboration

Big Data Technology Overview		
Term	Description	See Also
	become available—basically, adjusting as you go along *From PMBOK*, "Continuously improving and detailing a plan as more detailed and specific information and more accurate estimates become available as the project progresses, and thereby producing more accurate and complete plans that result from the successive iterations of the planning process".	
R	A free software programming language and software environment for statistical computing and graphics.	• R Project • R Studio • Wikipedia R Programming Language
Python	Often used as a scripting language, it is a general purpose, high-level programming language.	• Python • Wikipedia Python
Rolling wave planning	Rolling wave planning is a type of progressive elaboration planning. From PMBOK In this technique project management team plans tasks for the near future (few next iterations) as detailed as possible, while the work far in the future remains planned on a high level.	• Wikipedia Rolling Wave Planning
Scoop	A tool designed for transferring bulk data between Apache Hadoop and structured data stores such as relational databases or data warehouses.	• Apache Sqoop
Sentiment analysis	Extracting opinion or subjective information from source material such as social media.	• Wikipedia Sentiment Analysis
Storm	A distributed real-time computation system for processing fast, large streams of data. It is language independent. When used with Kafka, you can conduct reliable stream processing at a linear scale, assured that every message gets processed in real-time.	• Apache Storm
Stream Computing	Pulls in streams of data, processes the data and streams it back out as a single flow. Stream computing uses software algorithms that analyze the data in real time as it	• Wikipedia Stream Computing • IBM DataMag - The Role of Stream Computing in Big Data Architectures

Big Data Technology Overview		
Term	Description	See Also
	streams in to increase speed and accuracy when dealing with data handling and analysis.	
Tableau	BI tool that interface to data using SQL.	• Tableau Software
Talend	An open source software vendor that provides data integration, data management, enterprise application integration and big data software and services.	• Talend
UIMA	Unstructured Information Management applications are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user.	• Apache UIMA
Zookeeper	Zoo Keeper is a centralized configuration service and naming registry for large distributed systems. It is a high-performance coordination service for distributed applications.	• Apache Zookeeper