**What are Three V’s of Big Data?**

* **Volume**: How much data
* **Velocity**: How fast data is processed
* **Variety**: The various types of data

**Advantages of using Big Data platforms:**

Huge amount of data can be stored and processed quickly and cheaply, using a cluster of commodity hardware in a highly available and fault tolerant environment.

**Can 100 GB data be Big Data?**

If 100 GB data falls under 3v’s of Big Data, it may fall in Big data and may need Big data platforms to process it. However, people’s view may differ on this, but we should not forget that Big Data platform uses commodity grade hardware for low cost and high speed processing.

**How much data is Big Data?**

To be on safer side 1 TB of data can be called Big Data when apart from 1st V volume it also has 2nd V Velocity i.e. need to process data every now and then and the 3rd V means variety of unstructured data (Video, Audio, Image, Sensor-data, Documents etc.)

**What are different data formats in Big Data?**  
Big Data consists of following 3 categories of data.

* Structured Data: DB, DWH, ERP, CRM etc.
* Semi Structured Data: CSV, XML, JSON etc.
* Un Structured Data: Video, Audio, Image, Sensor Data, Documents etc.

**What is Data Lake in relation to Big Data?**A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi structured, and unstructured data. The data structure and requirements are not defined until the data is needed.

**What are few sectors using Big Data effectively?**

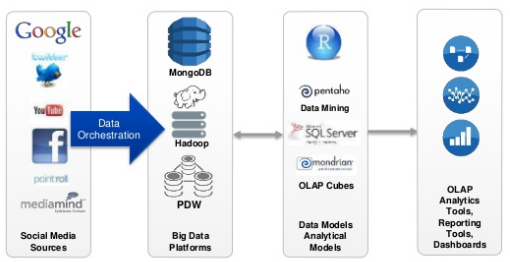
* **Banking and Securities:** Fraudulent transactions, Compliance and Customer segmentation etc.
* **Communication and Media:** Simultaneous real time report of several platforms.
* **Healthcare:** Collecting public health report, Identifying global spread of viruses etc.
* **Education:** To update and upgrade prescribed literature for rapid growth.
* **Manufacturing:** To enhance supply chain management
* **Insurance:** Developing new products based on user data, Risk Management.
* **Apart from above** Consumer Trade, Transportation, Energy and Sports are few other sectors using Big Data.

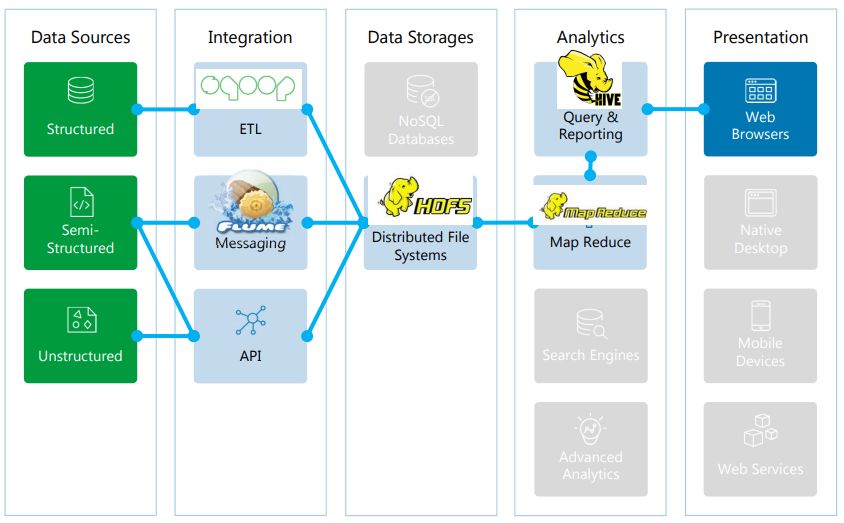
**What are some Technology companies using Big Data?** Fitbit, Smart meters by Energy Companies.

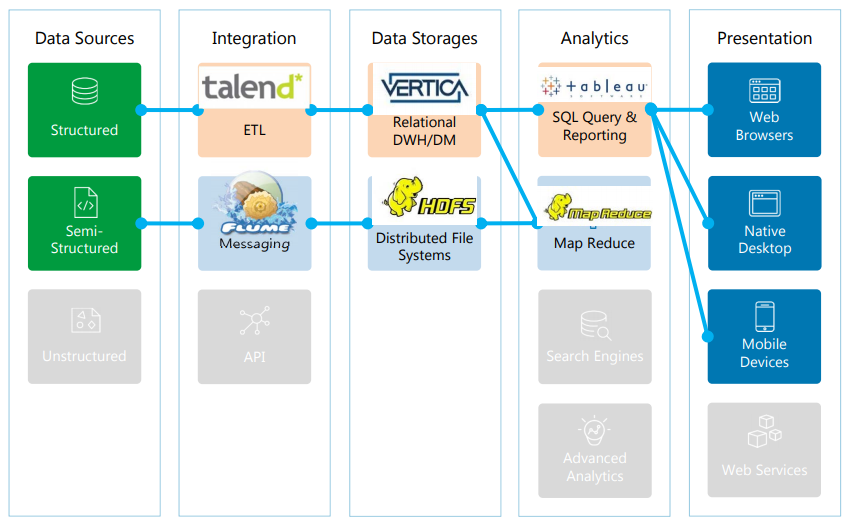
**How is Big Data changing the world?**

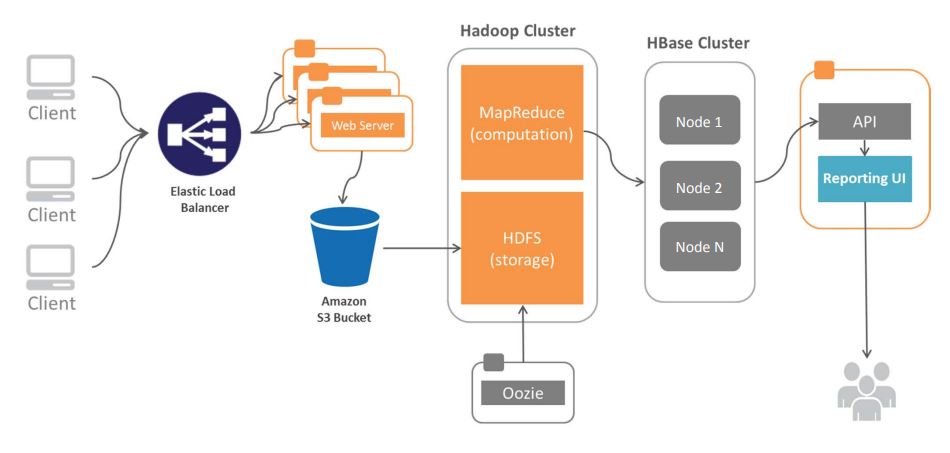
Big Data is being used by following departments. Security and Law enforcement, Science and research organizations. Retail Industry such as ecommerce and Finance such as stock trading. Big data is able to derive more value in these areas.

**Some reference Architecture for Big Data Project.**









**What are various Big Data Tools & Frameworks?**

**Hadoop**: Data Storage & Processing (HDFS+YARN)

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

**Spark**: Speed processing.

Apache Spark™ is a unified analytics engine for large-scale data processing. Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.

**HBase**: NoSQL database.

HBase is an open-source, non-relational, distributed database and runs on top of HDFS, providing Bigtable-like capabilities for Hadoop. Exposing data to transaction capability.

**Hive**: SQL type interface over HDFS.

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

**Sqoop**: HDFS to RDBMS and vice versa.

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

**Pig**: Similar to Hive, scripting language. Type pig on master node command prompt and grunt will start.

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark.

**Kafka**: Data streaming & processing.

Apache Kafka is an open-source stream-processing software platform developed by the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds.

**Flume:** Moving Large amount of data (weblogs).

A distributed, and reliable service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple & flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

**Oozie:** Workflow scheduler for jobs.

A workflow scheduler system to manage Hadoop jobs. Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions. Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability. Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs like Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop & Distcp as well as system specific jobs like Java programs and shell scripts.

**Ambari**: Web UI for managing Hadoop (<http://localhost:8888>, <http://127.0.0.1:8888>)

The Apache Ambari is web based UI for provisioning, managing, and monitoring Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management UI backed by RESTful APIs.

**Zookeeper**: co-ordination service to manage large set of hosts. ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

**Impala:** SQL query engine for Hadoop

Apache Impala is the open source, native analytic database for Apache Hadoop. Impala is shipped by Cloudera, MapR, Oracle, and Amazon. Impala provides low latency and high concurrency for BI/analytic queries on Hadoop (not delivered by batch frameworks such as Apache Hive).

**YARN:** Resource management job scheduling component YARN is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.

**Mesos:** Resource management and scheduling. Highly scalable alternative for YARN. It can also be implemented to work along with YARN.

**MapReduce:** Data processing in distributed and parallel.

**Tez:** Tez is an extensible framework for building high performance batch & interactive data processing applications, coordinated by YARN in Apache Hadoop. Hive through TEZ can often be faster than Hive through MapR

**Storm:** Processing streaming data Processing streaming data in real time. Spark Stream is another similar tool for treaming data.

Apache Drill, Hue, Phoenix, Presto, Zepplin etc.

**Apache Avro:** A data serialization system. [Link](https://www.youtube.com/watch?v=AU7S7pyk1qI).

Big Data with AWS: Intro with AWS stack