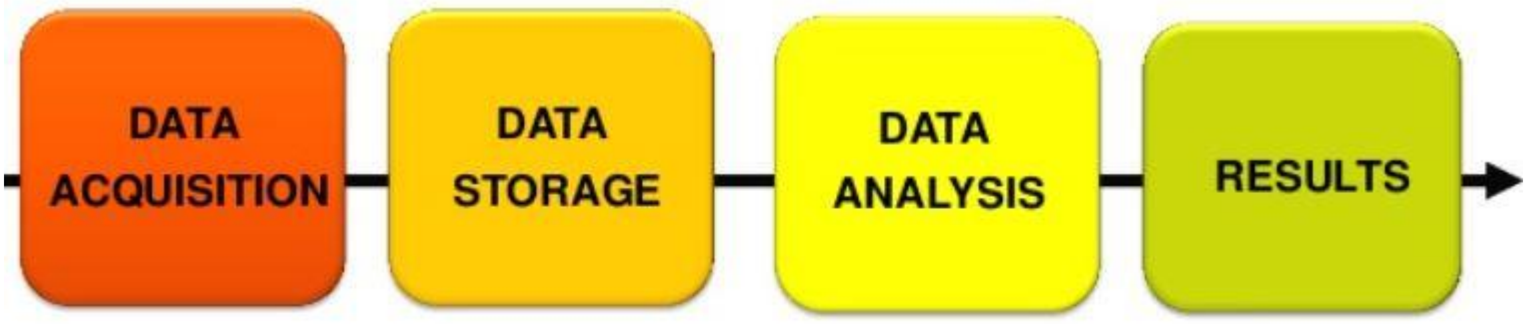
#TheStudyCircle

Your station for Big Data Technologies

Batch Vs Real time

Batch Processing	Real Time Processing
1. Large group of data/transactions is processed in a single run 2. Jobs runs without any manual intervention 3. The entire data is pre selected and fed using command line parameters and scripts 4. It is used to execute multiple operations, handle heavy data load, reporting and offline data workflow **Example :** Regular reports required decision making	1. Data processing takes place upon data entry or command receipt instantaneously. 2. It must execute on response time with stringent constraints.

Big Data processing pipeline

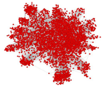


What is Hadoop?

1. It is a Big one of the popular big data processing framework.
2. It leverages the divide and conquer policy. The data is divided in multiple processing points(nodes) and the result is combined as one.
3. The framework is highly fault tolerant. The replication of data on cheap hardware is used for this purpose.

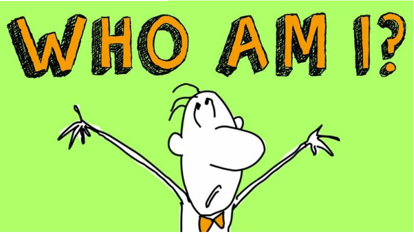
Developer uses “MAPREDUCE” **API** written in Java to manage **parallel data processing** in multiple nodes.
HDFS - Hadoop Distributed File System : Data storage system for Hadoop
It is mainly used for **Batch Processing.**

Limitation of Mapreduce in Hadoop

	Unsuitable for real time processing
	Unsuitable for trivial operations For operations such as filter and join the job needs to be re written which is a humongous task
	Unfit for large data on network It works on principle of data locality so works well for local data but unfit for large data on network
	Unsuitable for OLTP(Online transaction processing)
	Unfit for Graph processing
	Unfit for Iterative execution Being stateless mapreduce doesn’t fit with iterative processing

S

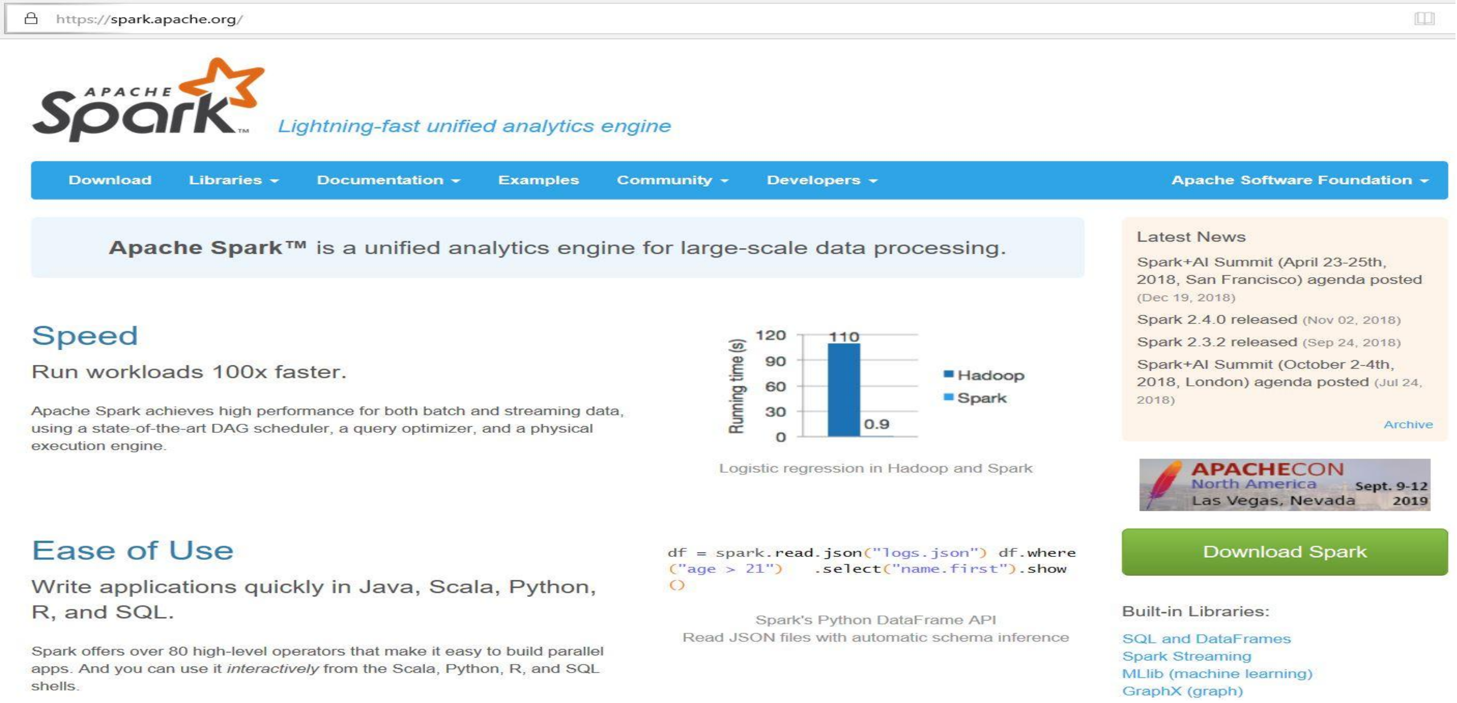
Spark



Apache Spark is the open standard for flexible in-memory data processing that enables batch, real-time, and advanced analytics on the Apache Hadoop platform

Know your product

https://spark.apache.org/



Lightning-fast unified analytics engine

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

```
df = spark.read.json("logs.json") df.where
("age > 21")     .select("name.first").show
()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Latest News

Spark+AI Summit (April 23-25th, 2018, San Francisco) agenda posted (Dec 19, 2018)

Spark 2.4.0 released (Nov 02, 2018)

Spark 2.3.2 released (Sep 24, 2018)

Spark+AI Summit (October 2-4th, 2018, London) agenda posted (Jul 24, 2018)

Archive

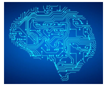
APACHECON North America Sept. 9-12 Las Vegas, Nevada 2019

Download Spark

Built-in Libraries:

- SQL and DataFrames
- Spark Streaming
- MLlib (machine learning)
- GraphX (graph)

Hadoop Ecosystem in Spark

	Batch Processing	Spark batch can be used over Hadoop Mapreduce
	Structured data analysis	Spark SQL can be used with SQL
	Machine learning Analysis	MLib can be used for clustering, recommendation and classification
	Interactive SQL Analysis	Spark SQL can be used over Impala
	Real time streaming Analysis	Spark streaming can be used

Use cases of real time analytics

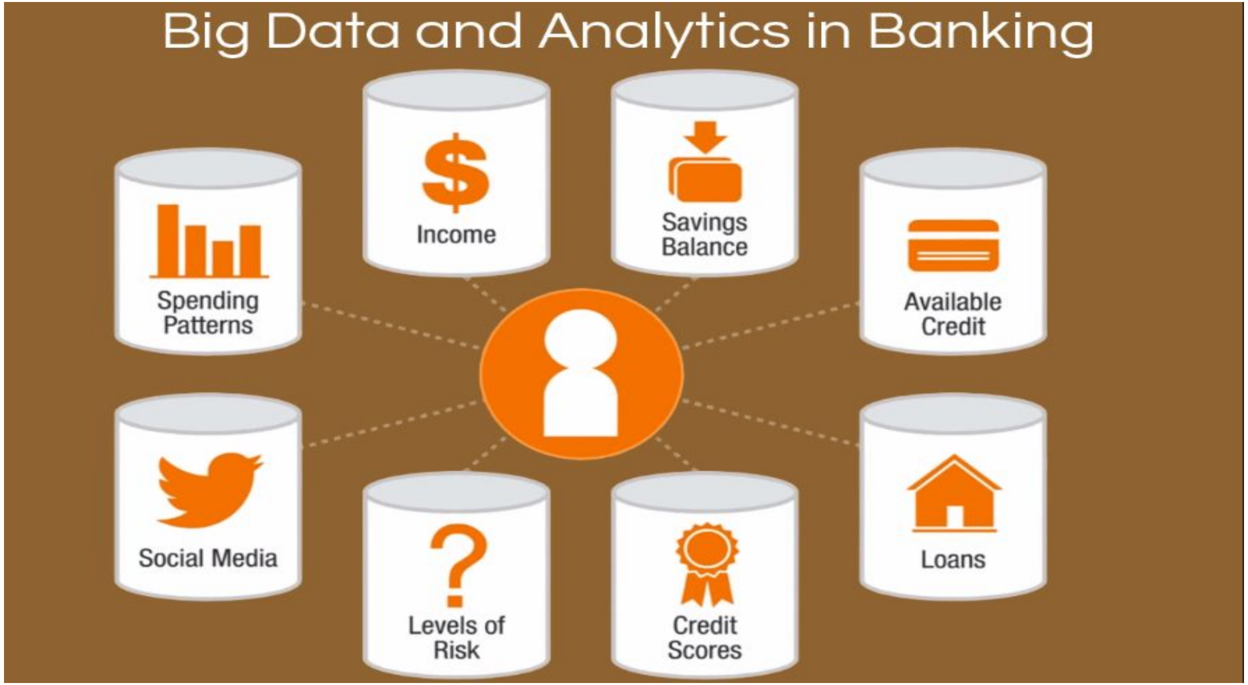
Banking

Government

Healthcare

Telecommunications

Banking Use Cases



Links

https://www.cloudera.com/products/open-source/apache-hadoop/apache-spark.html
https://www.bizofit.com/blog/analytics-banking-industry-important/
https://www.dezyre.com/article/how-is-hadoop-transforming-the-telecommunication-industry/88