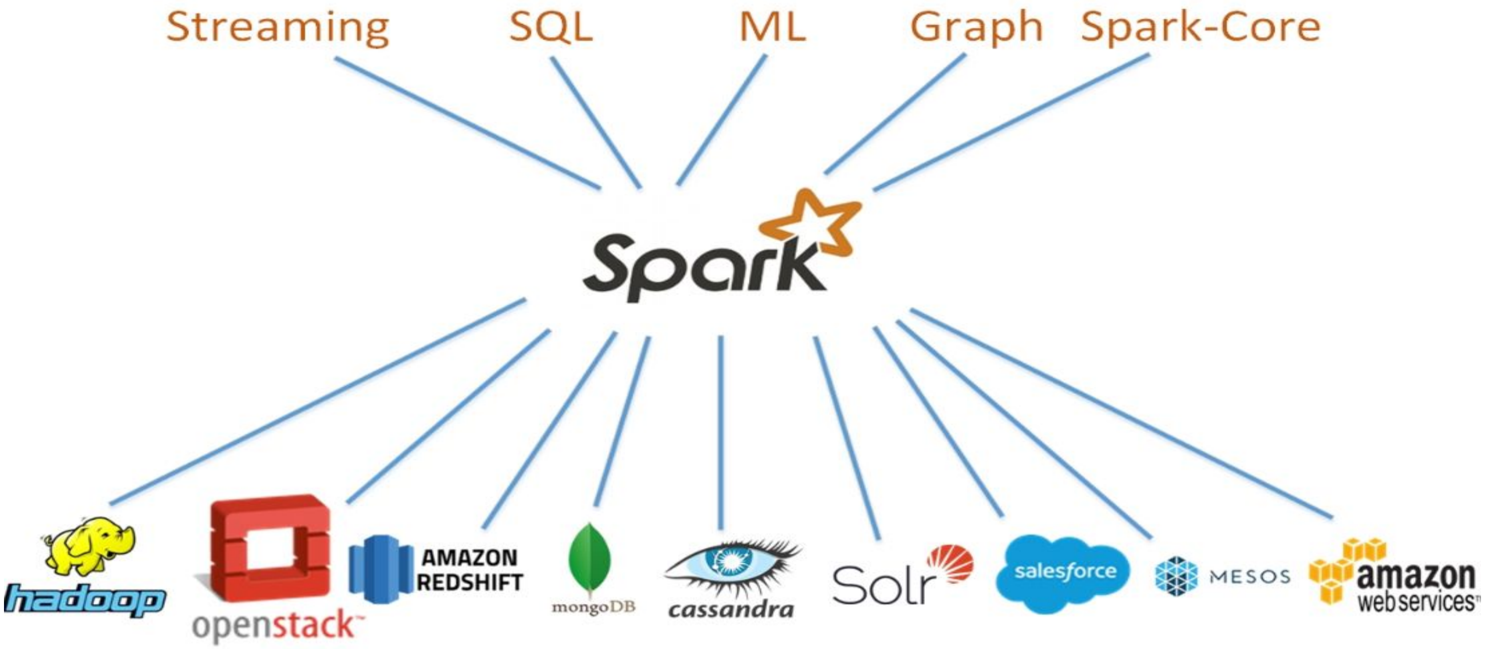
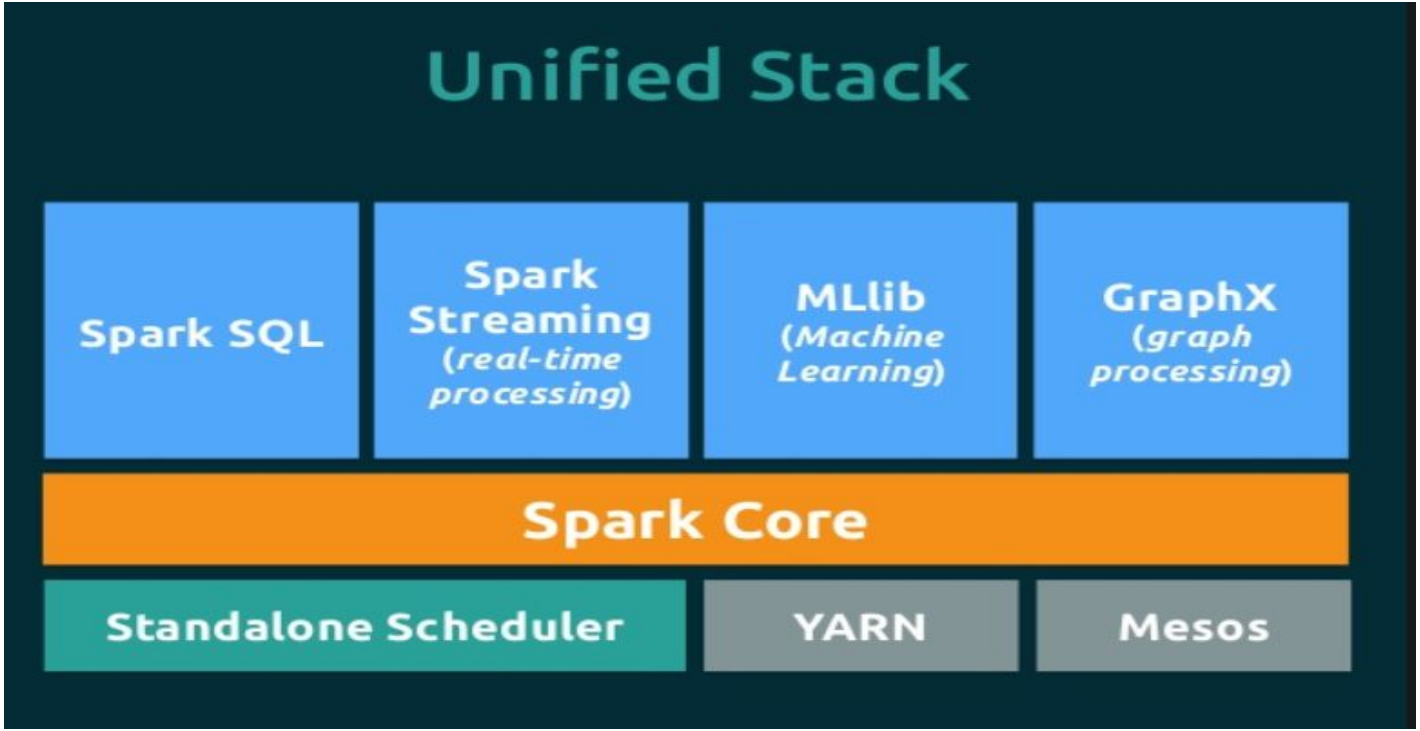
#TheStudyCircle

Your station for Big Data Technologies

Apache Spark - Unified Stack



Apache Spark



Unified Stack

Spark SQL

Spark Streaming (real-time processing)

MLlib (Machine Learning)

GraphX (graph processing)

Spark Core

Standalone Scheduler

YARN

Mesos

Apache Spark Core

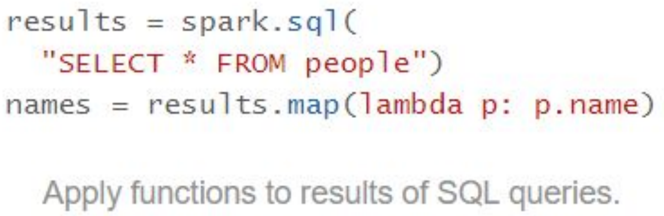
1. The core engine of Spark is commonly referred as spark-core. It is the foundation of spark architecture.
2. It provides following services
    a. Managing Memory pool
    b. Scheduling of tasks on Spark Cluster
    c. Recovering failed Jobs
    d. Providing support to work with wide variety of storages HDFS, S3, Cassandra etc.
3. Abstract users from low level technicalities of working on cluster.
4. Core has most important API : **RDD-API** - This is basis of higher level API.
5. Spark work as MPP (Massively parallel processing) system.

Business Analysts

Hive Impala

Hadoop Framework and MapReduce API posed challenge

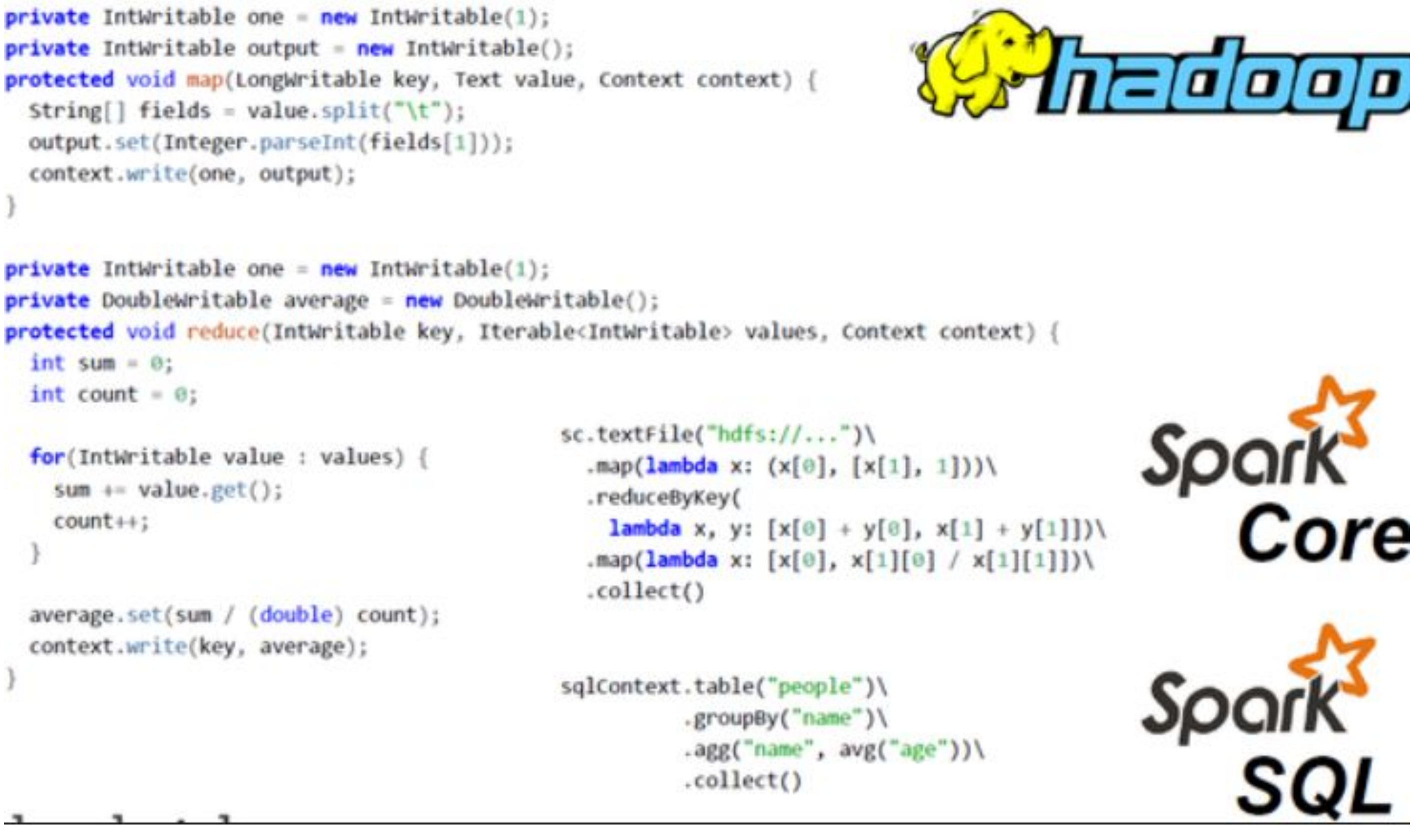
1. Business Analysts
2. People from database background



```
results = spark.sql(
  "SELECT * FROM people")
names = results.map(lambda p: p.name)
```

Apply functions to results of SQL queries.

Write Less Code



```java
private IntWritable one = new IntWritable(1);
private IntWritable output = new IntWritable();
protected void map(LongWritable key, Text value, Context context) {
  String[] fields = value.split("\t");
  output.set(Integer.parseInt(fields[1]));
  context.write(one, output);
}

private IntWritable one = new IntWritable(1);
private DoubleWritable average = new DoubleWritable();
protected void reduce(IntWritable key, Iterable<IntWritable> values, Context context) {
  int sum = 0;
  int count = 0;

  for(IntWritable value : values) {
    sum += value.get();
    count++;
  }

  average.set(sum / (double) count);
  context.write(key, average);
}
```

```python
sc.textFile("hdfs://...")\
  .map(lambda x: (x[0], [x[1], 1]))\
  .reduceByKey(
    lambda x, y: [x[0] + y[0], x[1] + y[1]])\
  .map(lambda x: [x[0], x[1][0] / x[1][1]])\
  .collect()
```

```python
sqlContext.table("people")\
          .groupBy("name")\
          .agg("name", avg("age"))\
          .collect()
```

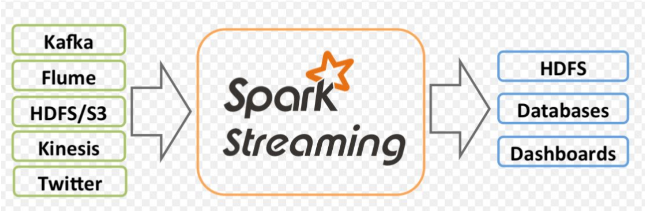
Apache Spark SQL

1. One of the most popular module designed for **structured and semistructured data processing.**
2. Query data inside Spark programs using SQL, DataFrame or Dataset API.
3. This feature is available in Java, Scala, Python and R.
4. Dataframe API allows access to : Hive, Avro, Paraquet, ORD, JSON and JDBC.
5. It has interconnectivity with Hive Metastore. Full compatibility with Hive data, Queries, UDF.
6. It has JDBC and ODBC interfaces for existing business tools.

Policy Bazaar use case

Retail store use case

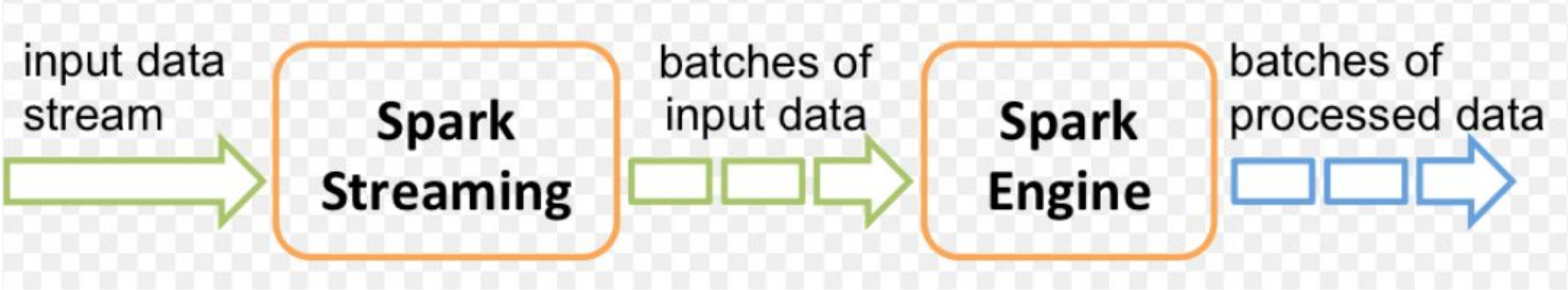
Apache Spark Streaming is the solution.



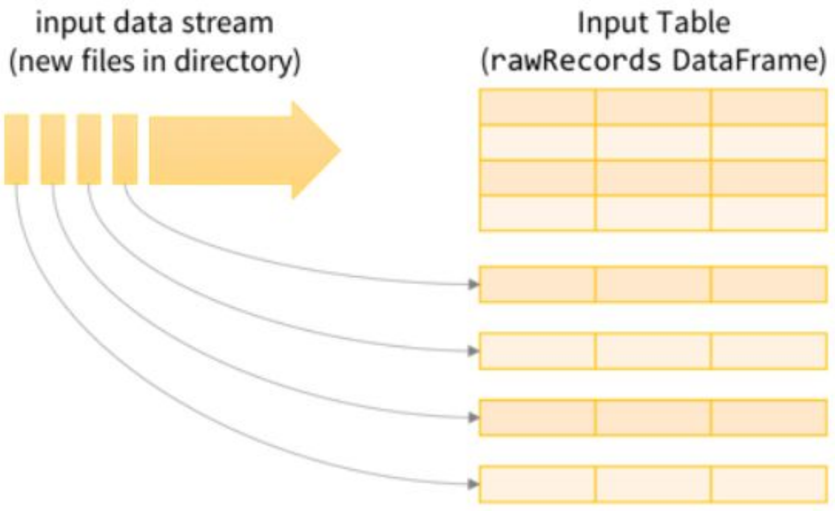
Banking use case

Apache Spark Streaming

1. Enables processing of data arriving in passive or live streams of data.
2. Passive streams - webserver logs, social media activity, twitter hashtag, sensor data from car/phone/house stored as file on cluster.
3. Spark-streaming has API similar to spark core which makes batch apps logic work for stream with small tweak. (Lambda architecture simplified)
4. A single API helps to create batch and stream app the only diff is batch apps has finite data in table and streams are infinite table getting continuously appended with data.



Structured streaming model - Spark



new records in data stream = new rows appended to unbounded Input Table

Structured Streaming Model
treat data streams as unbounded tables

Machine Learning

Machine learning is a branch of AI that gives computers ability to learn new patterns with no human intervention.

1. More Accurate results as more data is crunched
2. Roots in computation statistics
3. Used to predict fraud
4. Recommend new products and services to customers
5. Predict need of your car service

Machine learning is not new!!!!!!!!!!

It is top skill in **Gartner’s Hype Cycle**

Important questions

Why is machine learning an emerging technology? Why has it become so popular that it has made its way to **Gartner's Hype Cycle?**

1. Modern Data challenges are high dimensional. New techniques needed to solve problems
2. New computing paradigms, abundance of low cost high quality hardware and rich data sources
3. Need to outsmart competition
4. Move on from reporting past to predict the future as accurately as possible.
5. Limitation of traditional BI to limit decision on historical trends

Apache Spark MLLib

Machine learning library in spark designed to make ML scalable, approachable and easy for data scientists and engineers.

1. Iterative algorithms one of the key drivers behind creation of Spark.
2. Help built ML models from wide variety of data sources.
3. Model gets continuous training data from Spark streams.
4. Spark is superfast and 100 times faster than MAPReduce.
5. It has support for all popular algorithms : Logistic Regression, Support vector machines, classification and regression trees, random forest and gradient boosts, clustering, principal component analysis (PCA).

Key Features of Spark MLLib

ML Algorithms : Examples include classification, Regression, clustering and Collaborating filtering

Featurization : Feature extraction, transformation, dimensionality reduction and selection

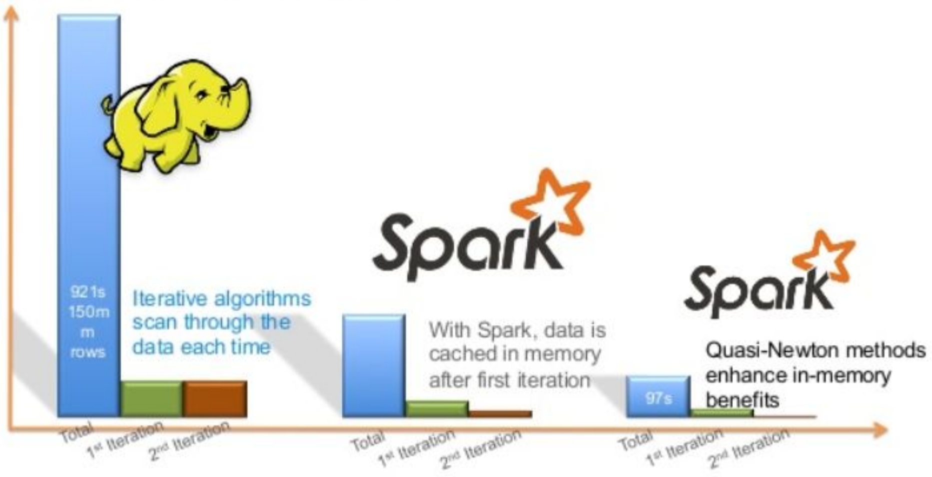
Pipelines : Tools for constructing, evaluating and tuning ML pipelines.

Persistence : saving and loading algorithms, models and pipelines

Utilities : Linear algebra, statistics and data handling

Credit Risk in Banks	
Self Driving Cars	
Sentiment Analysis	😍 🙂 😐 😠 😡
Fraud Detection	
Cyber security	

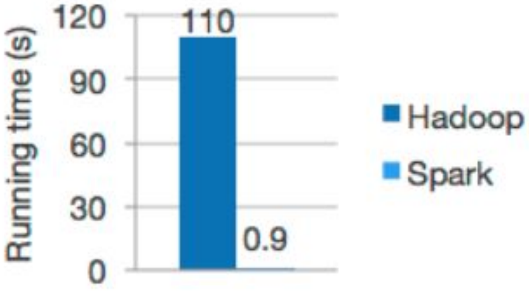
The Path to Innovation



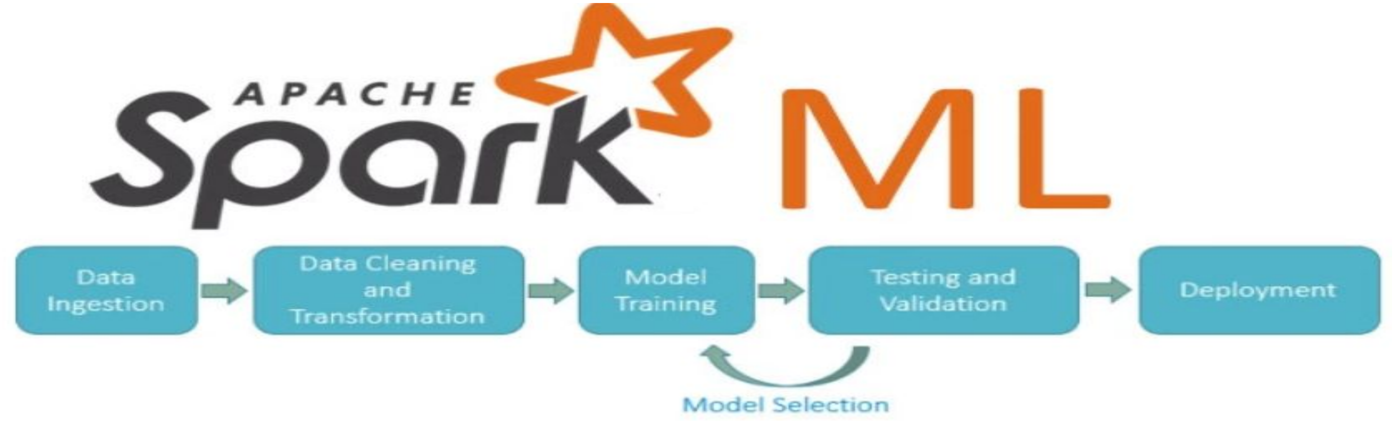
Iterative algorithms scan through the data each time

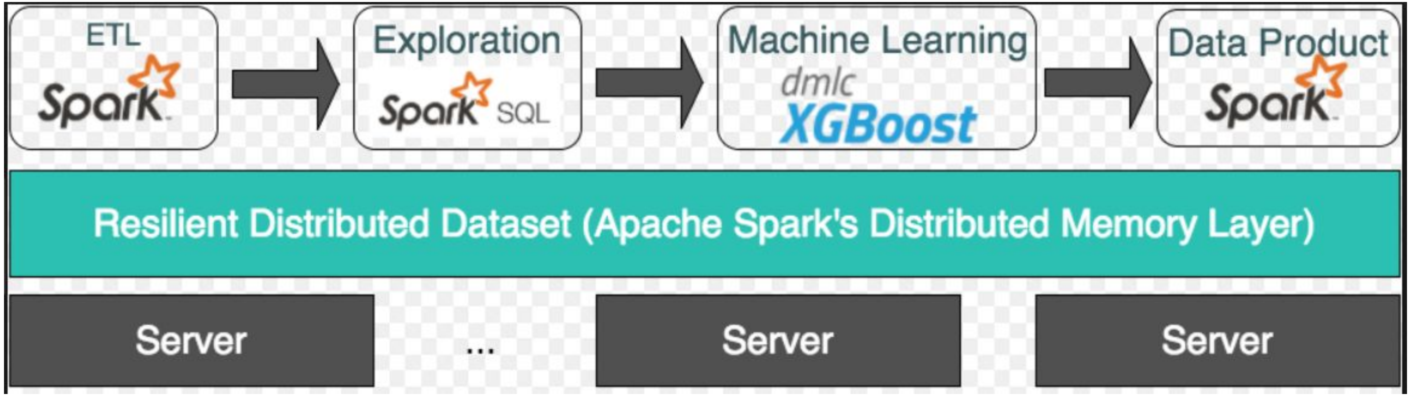
With Spark, data is cached in memory after first iteration

Quasi-Newton methods enhance in-memory benefits



Logistic regression in Hadoop and Spark


APACHE Spark ML
Data Ingestion
Data Cleaning and Transformation
Model Training
Testing and Validation
Deployment
Model Selection


ETL
Spark
Exploration
Spark SQL
Machine Learning
dmlc XGBoost
Data Product
Spark
Resilient Distributed Dataset (Apache Spark's Distributed Memory Layer)
Server
...
Server
Server

Hands on with Installation :

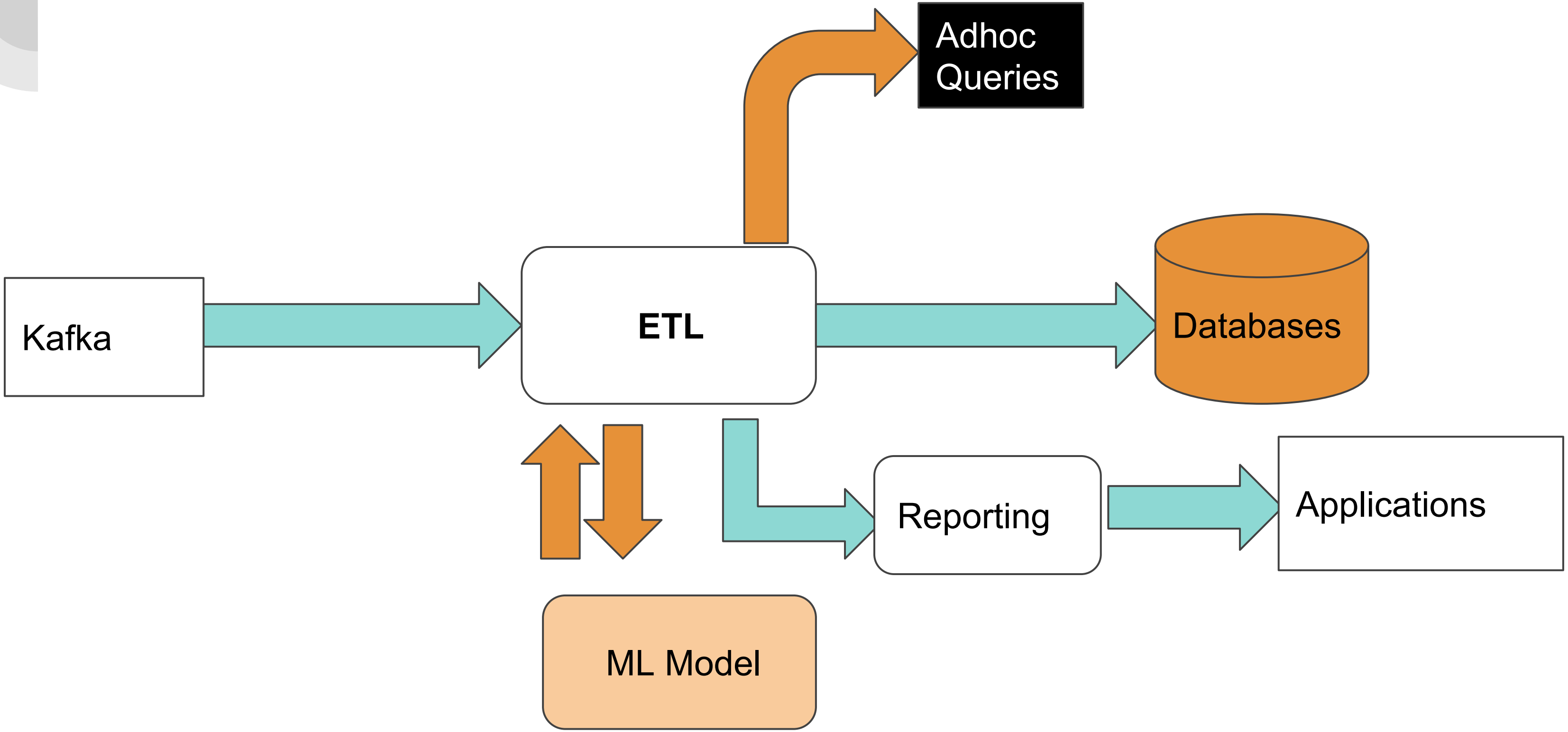
1. Amazon VM
2. Cloudera VM
3. Azure VM
4. Installation on Windows
5. Linux Installation - Install Spark latest version on that.

Steps for Spark installation on Linux :

https://github.com/wadhankarakash/TheStudyCircle/blob/master/SparkInstallation.pdf

Apache Spark Graph

1. API created to manipulate graphs.
2. The graphs and range from graph of web pages linked to each other via hyperlinks to a social network graph on twitter connected by followers or retweets or a facebook friend list.
3. Built in library for graph manipulation
4. Seamless work on both graphs and ETL, discovery analysis and iterative graph manipulation in single workflow.
5. Ability to combine transformations, machine learning and graph computations in single system
6. Ability to retain speed along with fault tolerance make it best for big data.
7. Built in algorithms : **PageRank, Connected components, Label propagation, SVD++ and triangular counter.**



Kafka

ETL

Adhoc Queries

Databases

Reporting

Applications

ML Model

https://tw.saowen.com/a/d145f98705c6c6472d7051c04f070498f5132419994aa2da71922a69ff78455e

https://www.dezyre.com/article/apache-spark-architecture-explained-in-detail/338