Apache Spark on Yarn

Haridas N

Agenda

- What’s apache spark
- How spark differs from standard map-reduce framework
- Spark on YARN
- Scala / Python Spark APIs
  - RDD, DataFrame, DataSet
  - Serialisation techniques
- Workshop:
  - Setup spark and configure it with our hadoop cluster
  - Submit some jobs

Images Need for this Session

```
docker pull haridasn/spark-2.4.0
docker pull haridasn/hadoop-2.8.5
docker pull haridasn/hadoop-cli
```

Tutorial Link: https://github.com/haridas/hadoop-env/blob/master/tutorials/spark-on-hadoop.adoc

Mapreduce framework

- Lot of disk operations.
- Well fit for batch processing over huge dataset.
- The APIs are bit low level, which leads to more work at application side.
- IO overhead is high.

Spark

- Does map-reduce in optimal way by using RAM as main storage.
- Optimised Memory operations.
- DAG based computation.
- Lazy evaluation and optimization of execution plan.

Spark Running modes

- Local Mode
  - Single JVM both driver and executor runs there. Good for local testing.
- Client Mode
  - Submitting jobs to spark cluster, The driver program runs on the client machine.
  - Easy way to see the program status.
- Cluster Mode
  - Driver program also runs on Cluster (One worker)
  - The client who submitted the job don’t need to be around.
  - Production grade jobs.

Spark APIs

- RDD
- DataSet
- DataFrame
- Transformer
- Serializers

RDD vs DataFrame vs DataSet

- RDD is the low level representation of data, Uses Java Serialisation by default.
- Immutable
- DataFrame and DataSet are typed RDDs
- Type information helps to do further optimization at low level (Network transfer, serialisation).
- Columnar Storage friendly.
- Apache Parquet and Apache Arrow project.

Speed up techniques

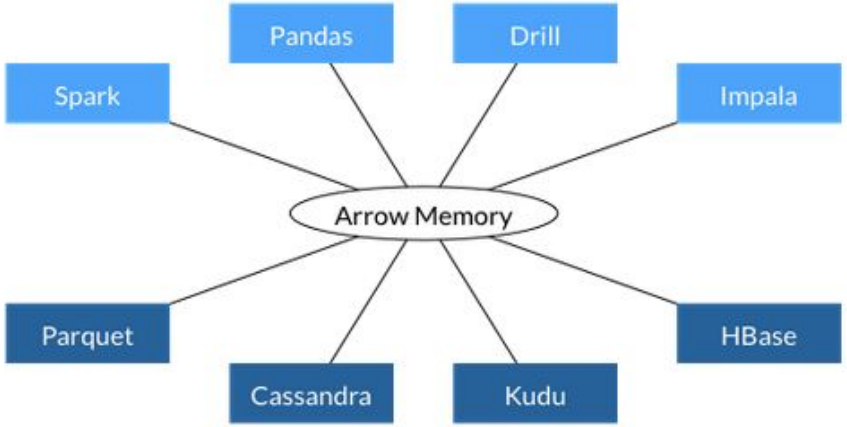
- Serialization
- Better memory representation.
- Columnar database

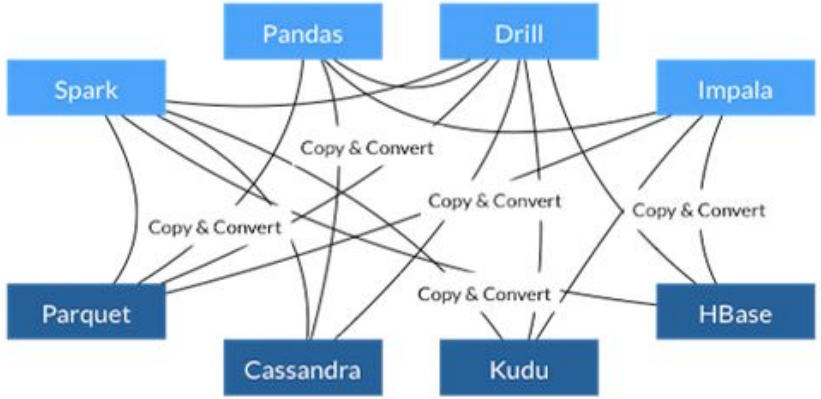
HDFS vs Object store

- Both aren’t POSIX compliant file system.
- HDFS more close to POSIX but some apis aren’t standard.
- Object stores are (s3, swift etc) purely a key-value blob store.
- Hadoop provides interfaces to interact with blob stores, which mimics the HDFS APIs over blob store.
- Spark or similar compute engines make use of this File System abstraction to interact with cloud storages.

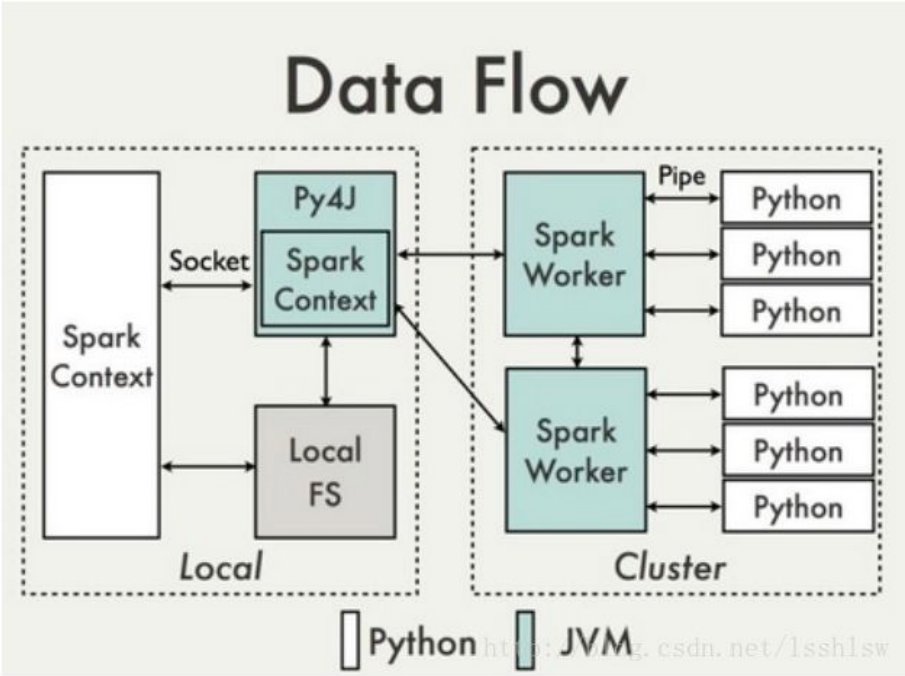
Columnar Storage

- How columnar storage change the big data performance scale
- Columnar databases, File storage formats, columnar memory representation.
- Parquet and Arrow project





Write spark jobs on python



Real workload - quick demo

- Document pre-processing
- Jobs are in python, using nltk and spacy NLP libraries.

Workshop

Thank you

Haridas N