SQOOP on SPARK

for Data Ingestion

Veena Basavaraj & Vinoth Chandar

@Uber

Currently @Uber on streaming systems.
@Cloudera on Ingestion for Hadoop.
@Linkedin on front-end service infra.

Currently @ Uber focussed on building a real time pipeline for ingestion to Hadoop. @Linkedin lead on Voldemort. In the past, worked on log based replication, HPC and stream processing.

Agenda

- Sqoop for Data Ingestion
- Why Sqoop on Spark?
- Sqoop Jobs on Spark
- Insights & Next Steps

Sqoop Before

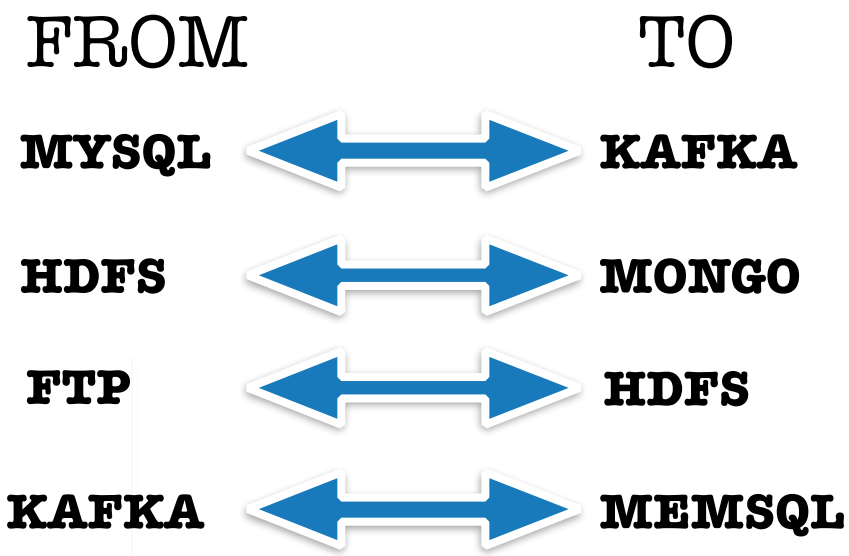
SQL  **HADOOP**

Data Ingestion

- Data Ingestion needs evolved
  - Non SQL like data sources
  - Messaging Systems as data sources
  - Multi-stage pipeline

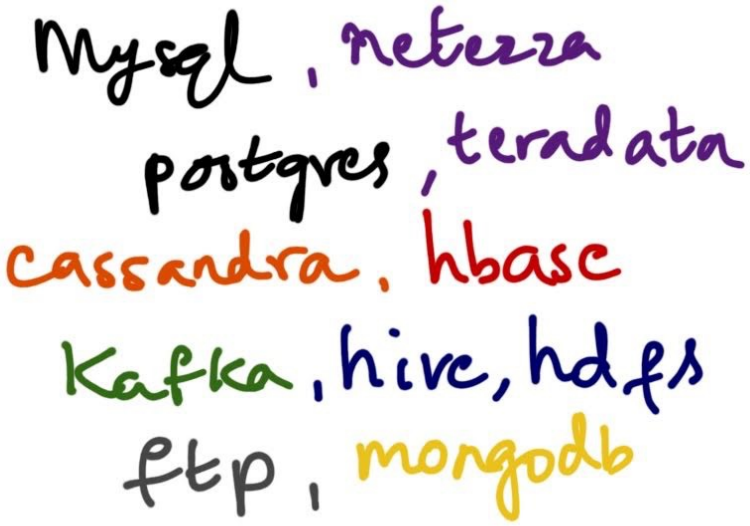
Sqoop Now

- Generic data Transfer Service
  - FROM ANY source
  - TO ANY target

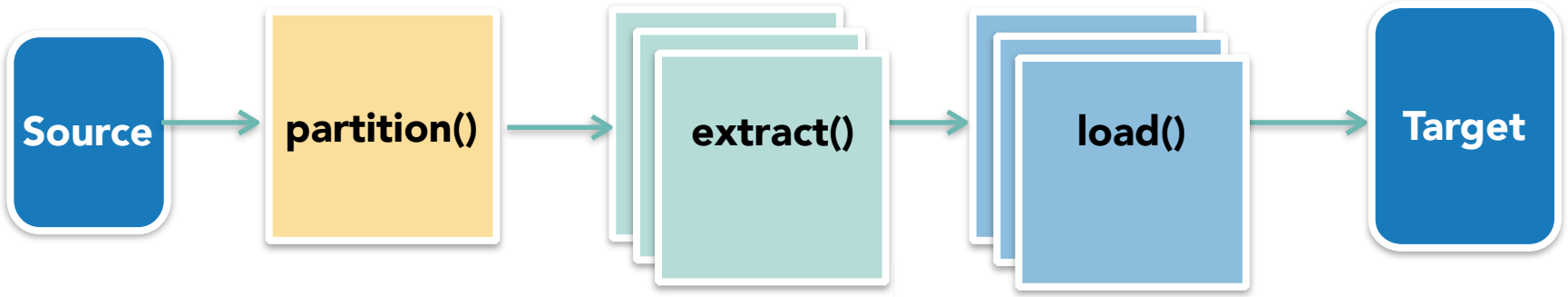


Sqoop How?

- **Connectors** represent Pluggable Data Sources
- Connectors are **configurable**
  - LINK configs
  - JOB configs



Sqoop Connector API



***No Transform (T) stage yet!*

Agenda

- Sqoop for Data Ingestion
- **Why Sqoop on Spark?**
- Sqoop Jobs on Spark
- Insights & Next Steps

It turns out...

- MapReduce is slow!
- We need Connector APIs to support (T) transformations, not just EL
- **Good news**! - Execution Engine is also pluggable

Why Apache Spark ?

- **Why not** ? ETL expressed as Spark jobs
- Faster than MapReduce
- Growing Community embracing Apache Spark

Why Not Use Spark Data Sources?

Sure we can ! but …

Why Not Spark DataSources ?

- Recent addition for data sources!
- Run MR Sqoop jobs on Spark with simple config change
- Leverage incremental EL & job management within Sqoop

Agenda

- Sqoop for Data Ingestion
- Why Sqoop on Spark?
- **Sqoop Jobs on Spark**
- Insights & Next Steps

Sqoop on Spark

- Creating a Job
- Job Submission
- Job Execution

Sqoop Job API

- Create Sqoop Job
  - Create FROM and TO job configs
  - Create JOB associating FROM and TO configs
- SparkContext holds Sqoop Jobs
- Invoke SqoopSparkJob.execute(conf, context)

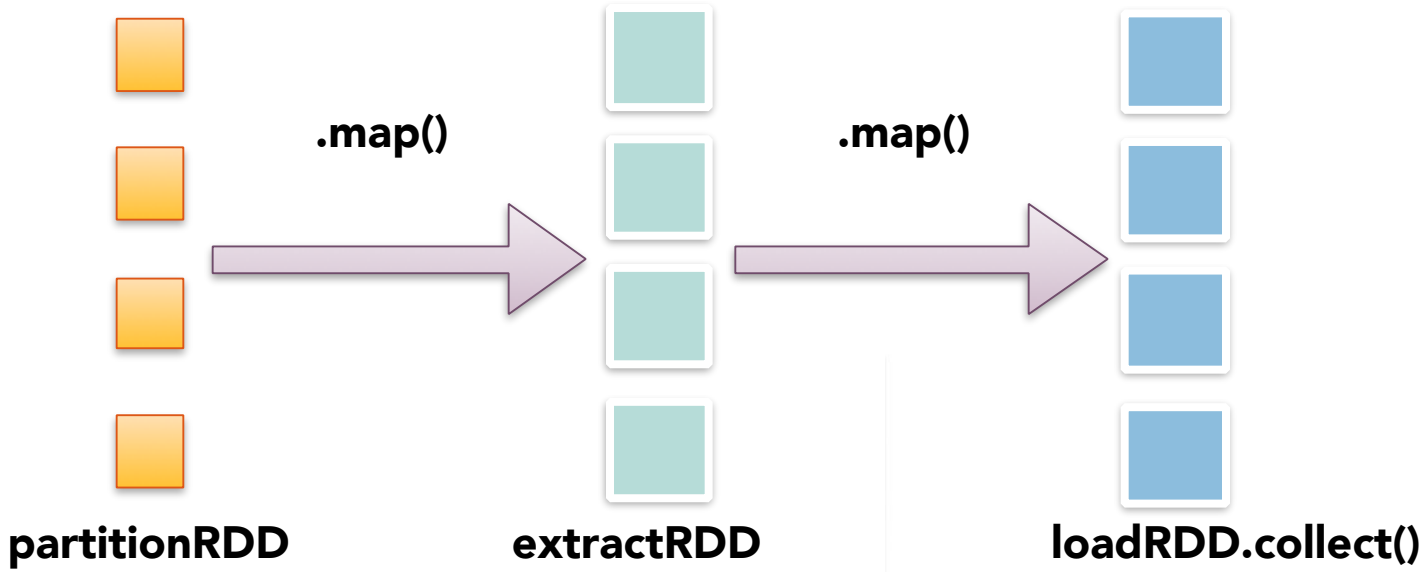
Spark Job Submission

- We explored a few options.!
  - Invoke Spark in process within the Sqoop Server to execute the job
  - Use Remote Spark Context used by Hive on Spark to submit
  - Sqoop Job as a driver for the Spark submit command

Spark Job Submission

- Build a “uber.jar” with the driver and all the sqoop dependencies
- Programmatically using Spark Yarn Client (non public) or directly via command line submit the driver program to yarn client/
  - **bin/spark-submit —class org.apache.sqoop.spark.SqoopJDBCHDFSJobDriver --master yarn /path/to/uber.jar —confDir /path/to/sqoop/server/conf/ —jdbcString jdbc://myhost:3306/test —u uber —p hadoop —outputDir hdfs://path/to/output —numE 4 —numL 4**

Spark Job Execution



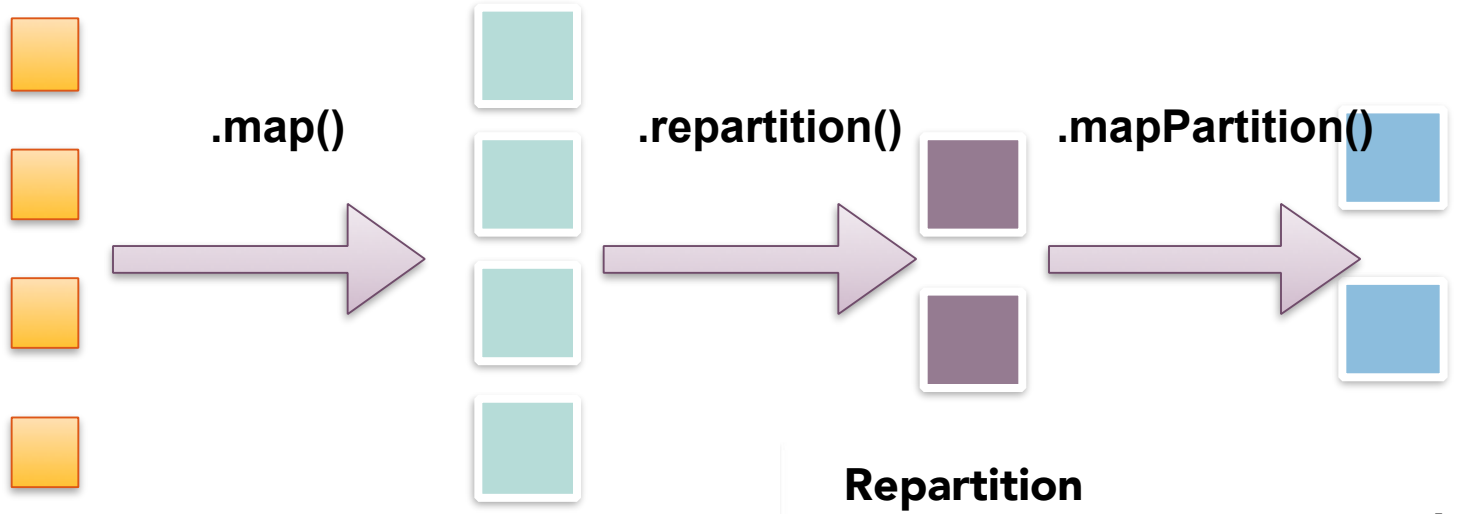
Spark Job Execution

```
SqoopSparkJob.execute(…)
1   List<Partition> sp = getPartitions(request,numMappers);
    JavaRDD<Partition> partitionRDD = sc.parallelize(sp, sp.size());

2   JavaRDD<List<IntermediateDataFormat<?>>> extractRDD =
    partitionRDD.map(new SqoopExtractFunction(request));

3   extractRDD.map(new SqoopLoadFunction(request)).collect();
```

Spark Job Execution



Compute Partitions

Extract from MySQL

Repartition to limit files on HDFS

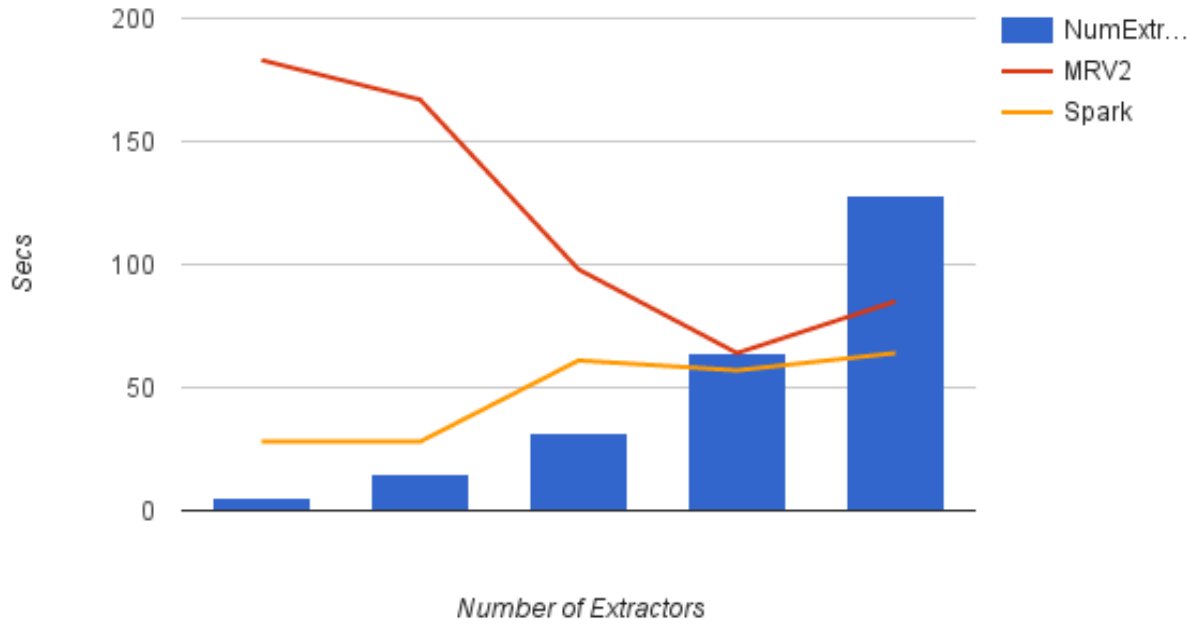
Load into HDFS

Agenda

- Sqoop for Data Ingestion
- Why Sqoop on Spark?
- Sqoop Jobs on Spark
- **Insights & Next Steps**

Micro Benchmark: MySQL to HDFS

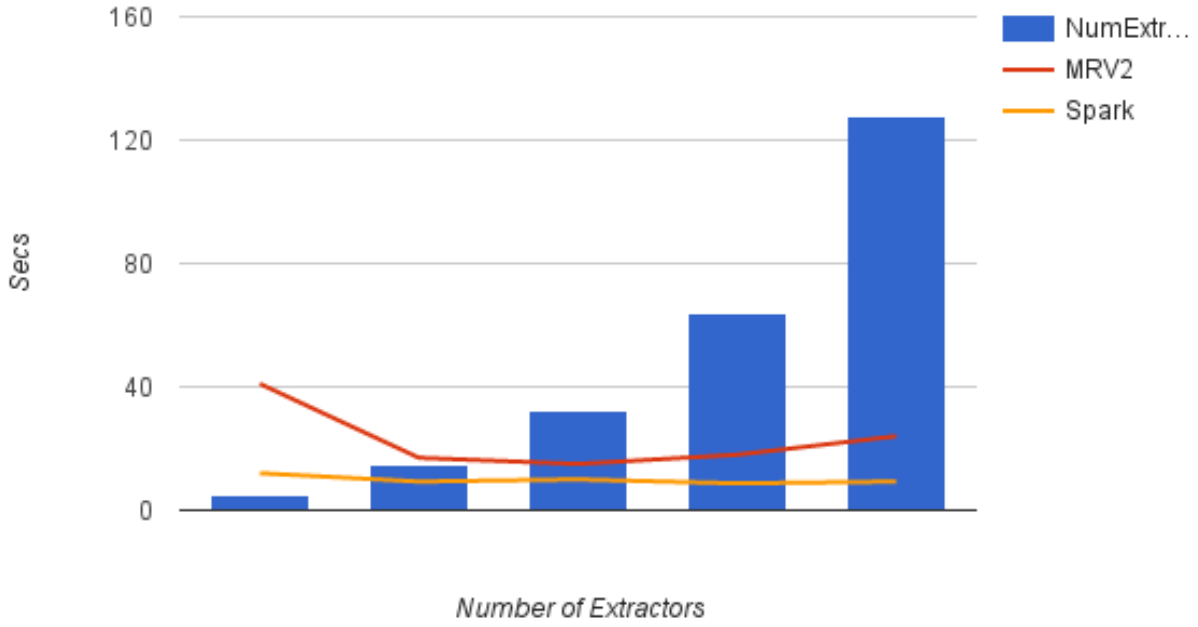
Table w/ 300K records, numExtractors = numLoaders



Micro Benchmark: MySQL to HDFS

Table w/ 2.8M records, numExtractors = numLoaders

good partitioning!!



What was Easy?

- **NO** changes to the Connector API required.
- Inbuilt support for Standalone and Yarn Cluster mode for quick end-end testing and faster iteration
- Scheduling Spark sqoop jobs via Oozie

What was not Easy?

- No clean public Spark Job Submit API. Using Yarn UI for Job status and health.
- Bunch of Sqoop core classes such as IDF had to be made serializable
- Managing Hadoop and spark dependencies together in Sqoop caused some pain

Next Steps!

- Explore alternative ways for Spark Sqoop Job Submission with Spark 1.4 additions
- Connector Filter API (filter, data masking)
- SQOOP-1532
  - https://github.com/vybs/sqoop-on-spark

Sqoop Connector ETL

***With Transform (T) stage!*

Questions!

- Thanks to the Folks @Cloudera and @Uber !
- You can reach us @vybs, @byte_array