CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2016
Doc 28 Compiling Spark
Dec 13, 2016

How to Compile Spark Program - 1

You need three jar files

```
javac -cp /SparkInstall/jars/scala-library-2.11.8.jar: \
                        /SparkInstall/jars/spark-core_2.11-2.0.1.jar: \
                        /SparkInstall/jars/spark-sql_2.11-2.0.1.jar:.
SampleProgram.java
```

How to Compile Spark Program - 2

Use an IDE

How to Compile Spark Program - 3

Use Maven

Logging

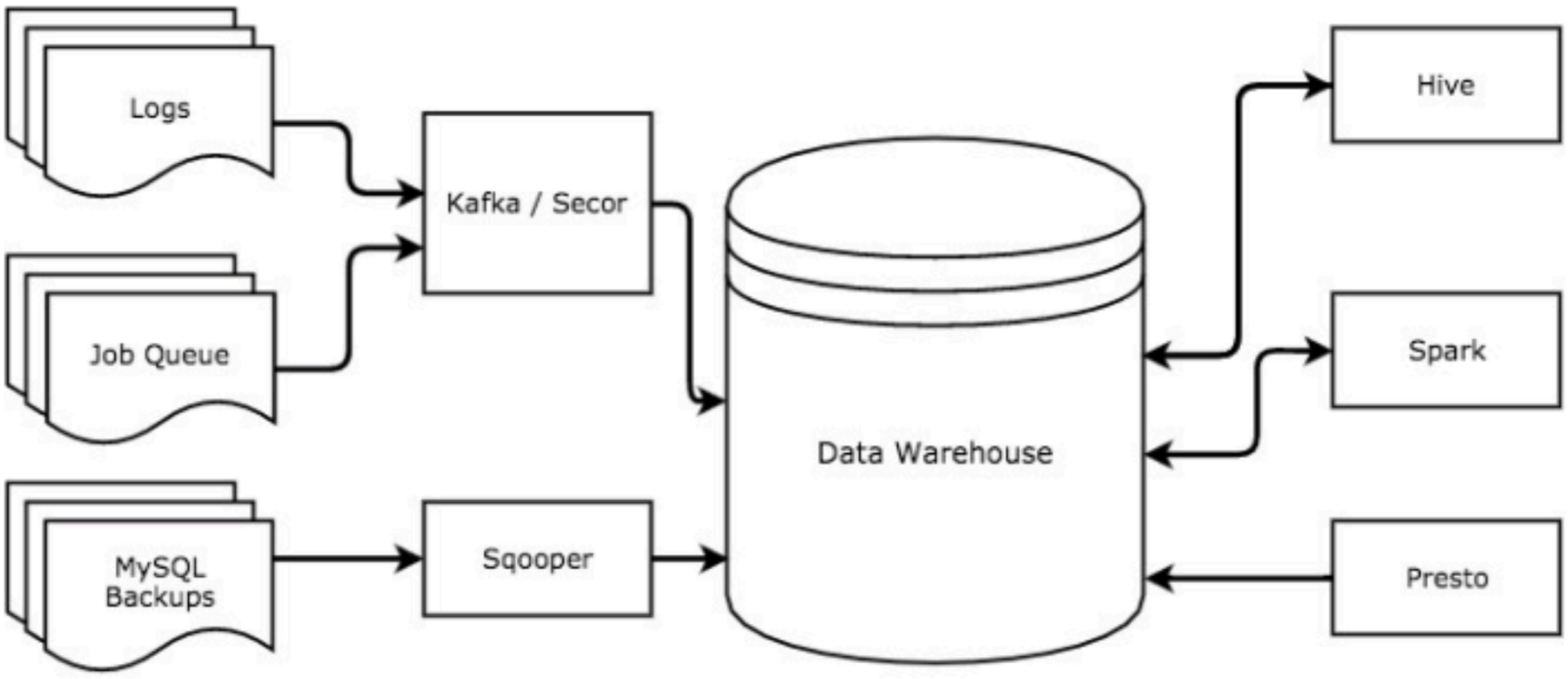
```
import org.apache.log4j.LogManager;
import org.apache.log4j.Logger;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.api.java.JavaRDD;
import java.util.Arrays;

public final class SampleProgram {
    public static void main(String[] args) throws Exception {

        Logger log = LogManager.getRootLogger();
        log.warn("start");
        JavaSparkContext sc = new JavaSparkContext();
        JavaRDD<Integer> rdd = sc.parallelize(Arrays.asList(1, 2, 3, 4, 5, 6, 7,8));
        rdd.saveAsTextFile("outputDir");
        log.warn("*********end");
        sc.stop();
    }
}
```

Data Wrangling at Slack

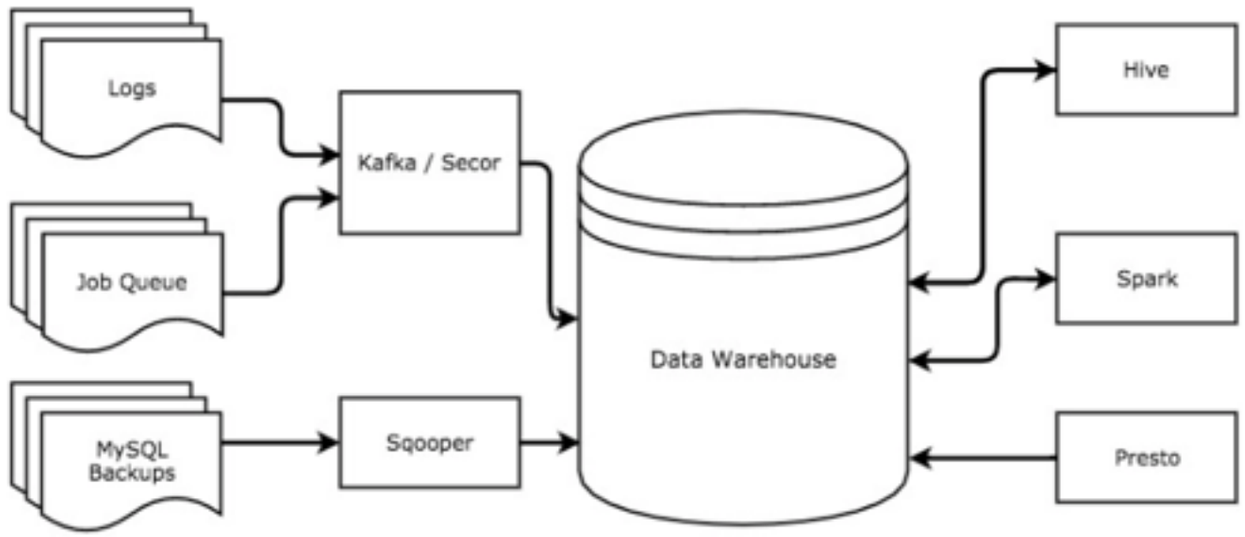
https://slack.engineering/data-wrangling-at-slack-f2e0ff633b69#.dfrmdy6b5



Sqooper

In house

Imports MySQL into S3



Kafka

Scalable append message log

Secor

Persists data from Kafka into S3

Hive

Converts SQL-like queries into MapReduce

Presto

Distributed SQL query engine

Interactive queries

Spark

Use Scala

Tying It All Together

How to make sure the tools can work together

- Data formats, transport layers, protocols

Apache Thrift

- Cross-language services development
- C++, Java, Python, PHP, Erlang, Perl, Haskell, C#, JavaScript, Node.js
  - Smalltalk, OCamal, Delphi

Common type system

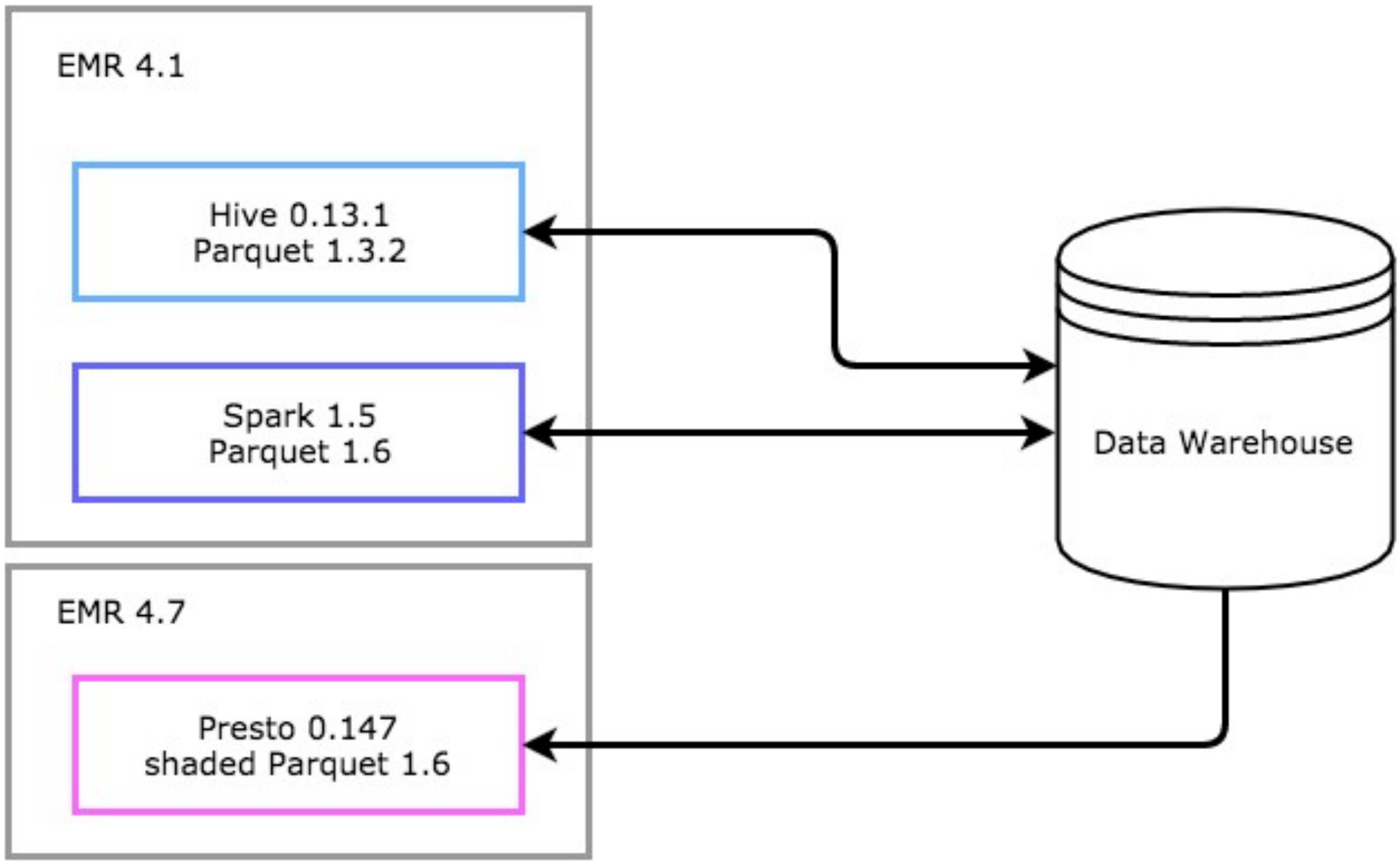
Transport - handle converting native types to/from common type system

Tying It All Together

- Thrift
  - Common typed schema so have structured data
- Parquet
  - Column storage format
  - Available in all Hadoop ecosystem tools
- Hive Metastore
  - Data abstraction
    - How do tools import, export, encode, decode data
  - Used as ground truth for data & schema

Problems

Each tool has different version of Parquet
Each has different set of bugs



Absence of Data - null

Hive 0.13 with Parquet

null throws an exception

Null as key in a map

Some versions of Parquet handle this other throw an exception

Schema Evolution

Had three different schemas for data and file formats

Any change to schema means

- All three have to change
- Update all old data