Processing Web Server Logs

Faculty: Michael Enudi

About Me.

- Lives and works in Johannesburg, South Africa
- Senior Software engineer with over 10 years of working experience writing enterprise java applications, architecting data solutions.
- Cloudera Certified Spark and Hadoop Dev.
- Oracle Certified SQL Expert
- Oracle Certified Java Master
- Sun Certified Java Business Component Dev.
- Sun Certified Java Programmer
- Big data enthusiast

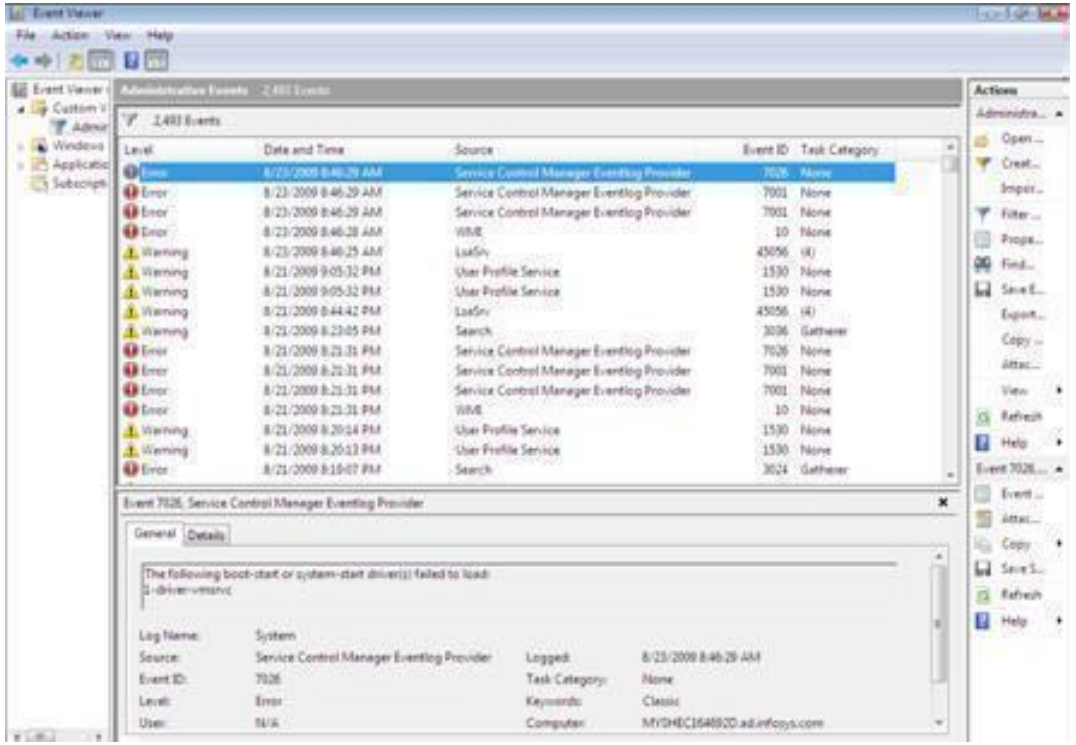
LinkedIn → https://www.linkedin.com/in/michaelenudi

What is a Log File

In computing, a logfile is a file that records either events that occur in an operating system or other software runs, or messages between different users of a communication software.[citation needed] Logging is the act of keeping a log. In the simplest case, messages are written to a single logfile.

Types of Log Files

- Proxy log
- Transaction log
- Event log
- Message log
- Application log
- Web server log



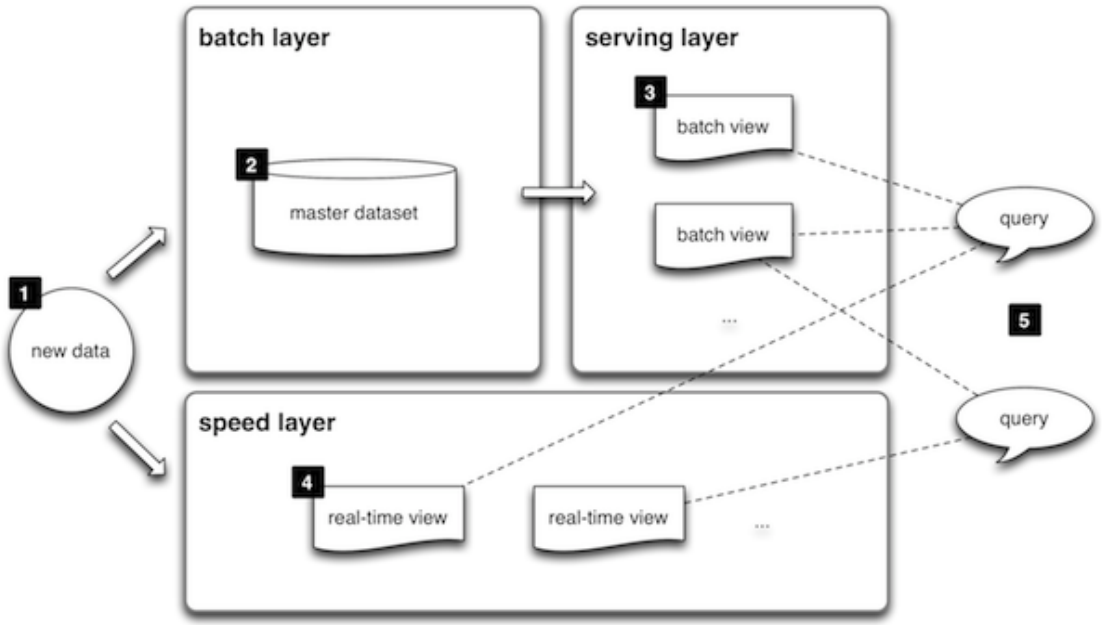
Web Server Log Processing Use Case

- Application Health Monitoring
- Fraud - Security
- User Pattern (sessionizing a click stream)
- User Experience
- Support Triage
- Metric Data Collection

Some Log File Processing Tools

- Flume
- Logstach (on Elastic search)
- Splunk

Lambda Architecture

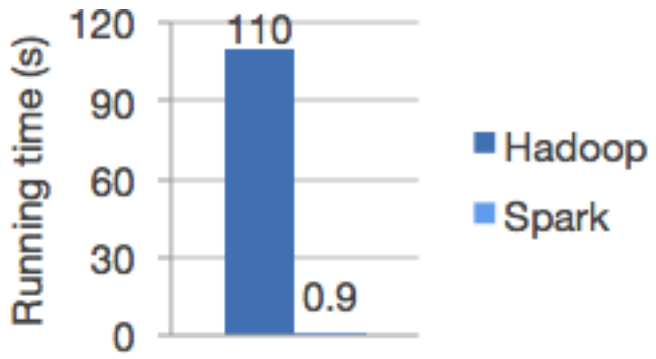


Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch- and stream-processing methods. (Wikipedia)

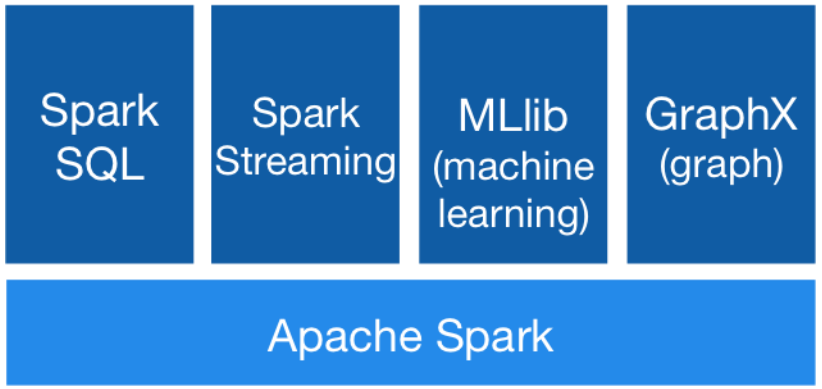
The approcah is to divide the entire data processing infrastructure into three layers:

- Batch Layer
- Speed Layer
- Serving Layer

Apache Spark



An engine for fast and efficient data processing. It is a natural success of Hadoop's MapReduce framework that comes with in-memory data structure (RDD), directed-acyclic graphs of steps in the job pipeline and multi-language platform to write data processing application on and beyond Hadoop.

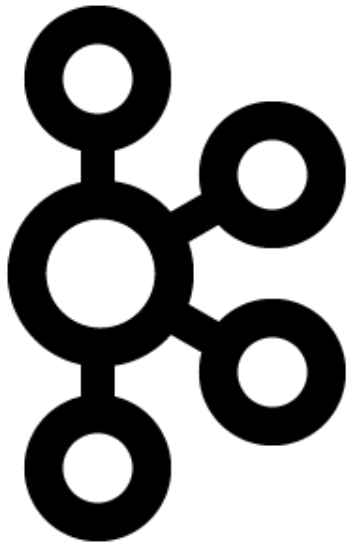


It is seriously becoming a sub-ecosystem as it runs other sub framework for data processing like Spark SQL, Spark Stream, Spark ML and many more.

Scala and its build tool (Sbt)

Scala is a general purpose language that mixes various paradigms and runs on the java virtual machine (JVM). With Scala, you can perfectly mix functional and reactive styles of programming with Object-oriented programming.

SBT or Scala build tool - the maven for Scala.



APACHE kafka™

A distributed streaming platform

Use case:

- Messaging
- Website Activity Tracking
- Metrics
- Log Aggregation
- Stream Processing
- Event Sourcing

Case study

Requirement

Given the two web server log dataset available for download from http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html , we must make analysis, event processing, and data retrieval of log data for varying kind possible.
Data from the log must be available either using low-latency querying tool or real-time event reporting and analysis.

Input Analysis

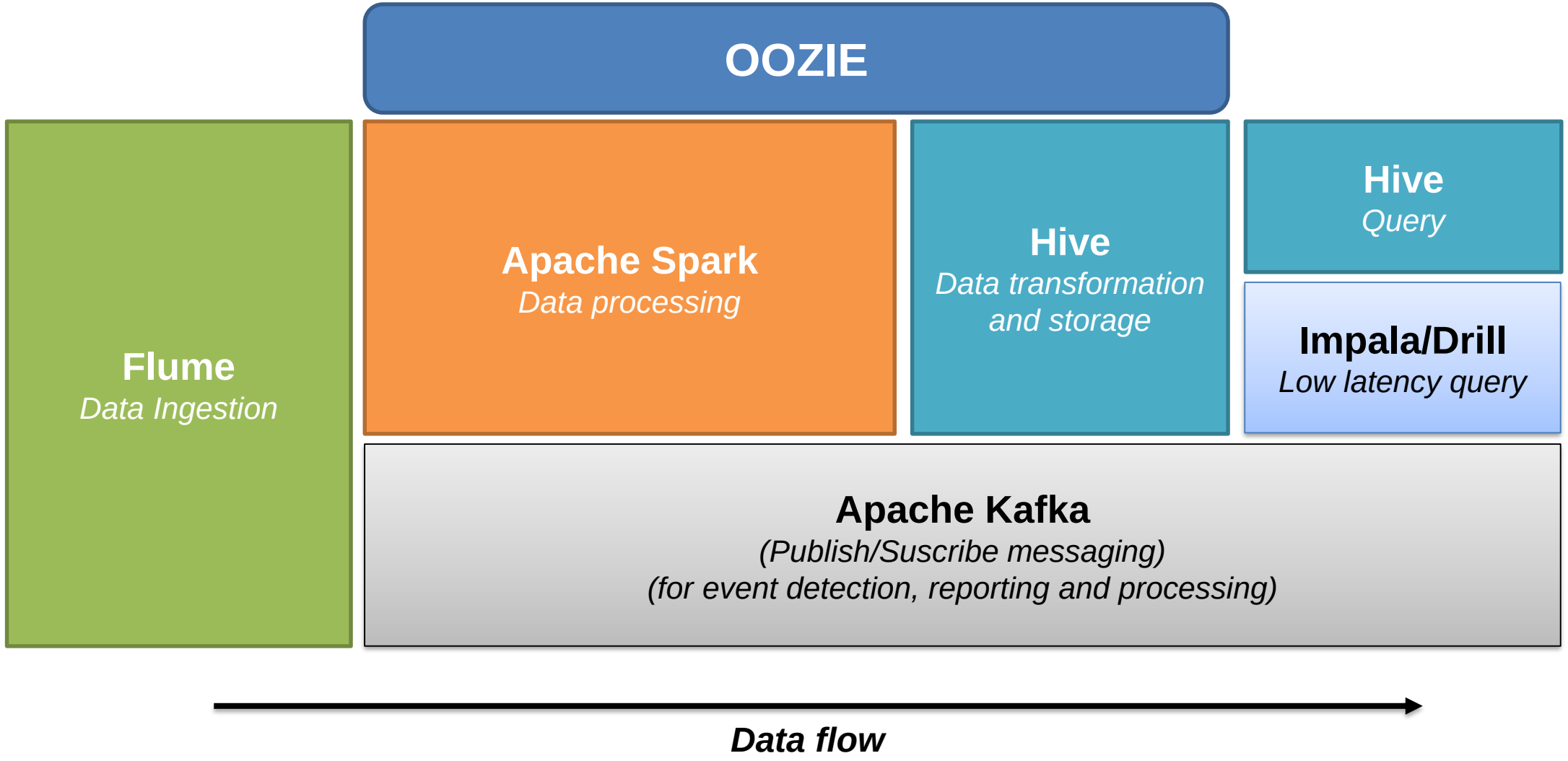
```
in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 1839
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 304 0
```

- host that made the request to the server
- user-identifier is the RFC 1413 identity of the client.
- User Id of the person requesting the document
- date and time when the call was made with time zone for the caller
- the request line for the call
- the http status code for the call
- the size in bytes of the response

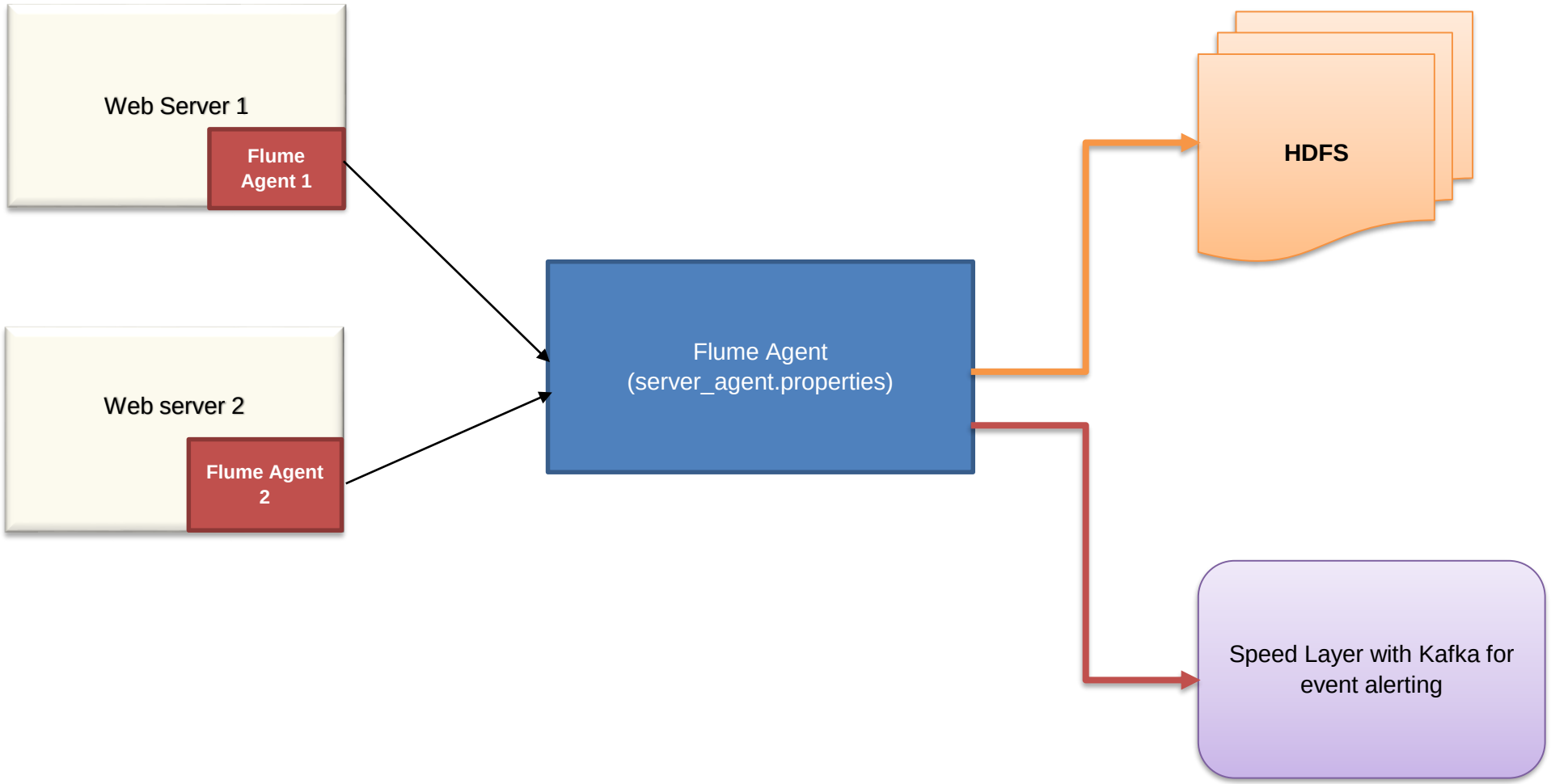
Some other log may contain information such as

- Referral url
- The user agent for the request.

Architecture



Architecture: Ingestion

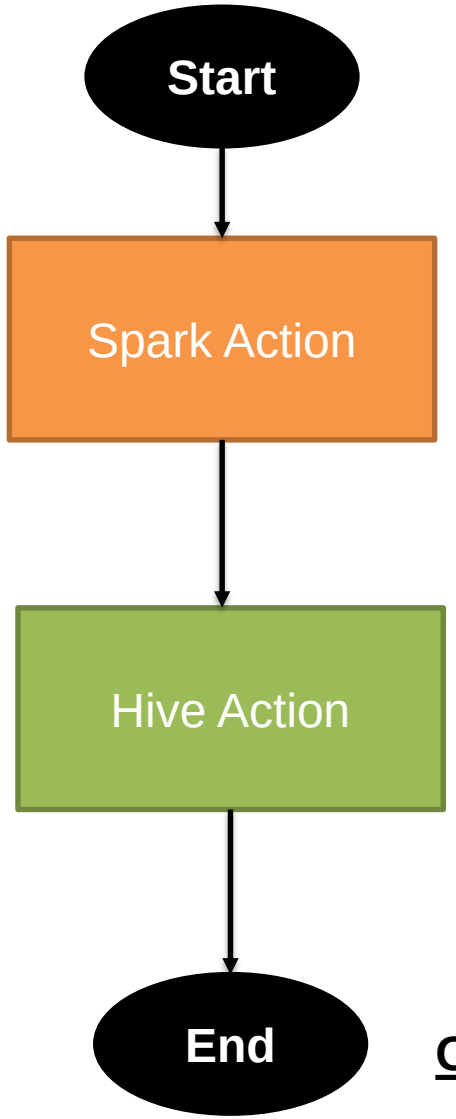


Architecture: Processing

A simple string-based transformation from log entry to a parquet data structure containing all fields

Using UDFs to further-parse fields to provide more granular values of fields from the spark transformation.

Architecture: Coordination



Oozie Workflow

Once every day.
Begin 12 am for data of the previous day

Oozie Coordinator

Requirements

- Cloudera Quickstart VM 5.7 or 5.8
- Scala SDK and Runtime
- Scala build tool

Thank You