Use Splunk With Big Data Repositories Like Spark, Solr, Hadoop And Nosql Storage

Raanan Dagan, May Long

Big Data Architect, Splunk

.conf2016

splunk>

Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Agenda

Use Cases:

- Fraud With Solr, Splunk, And Splunk Analytics For Hadoop
- Business Analytics With Cassandra, Splunk Cloud, And Splunk Analytics For Hadoop
- Document Classification With Spark And Splunk
- Network IT With Kafka And Splunk Kafka Add On
- Demo

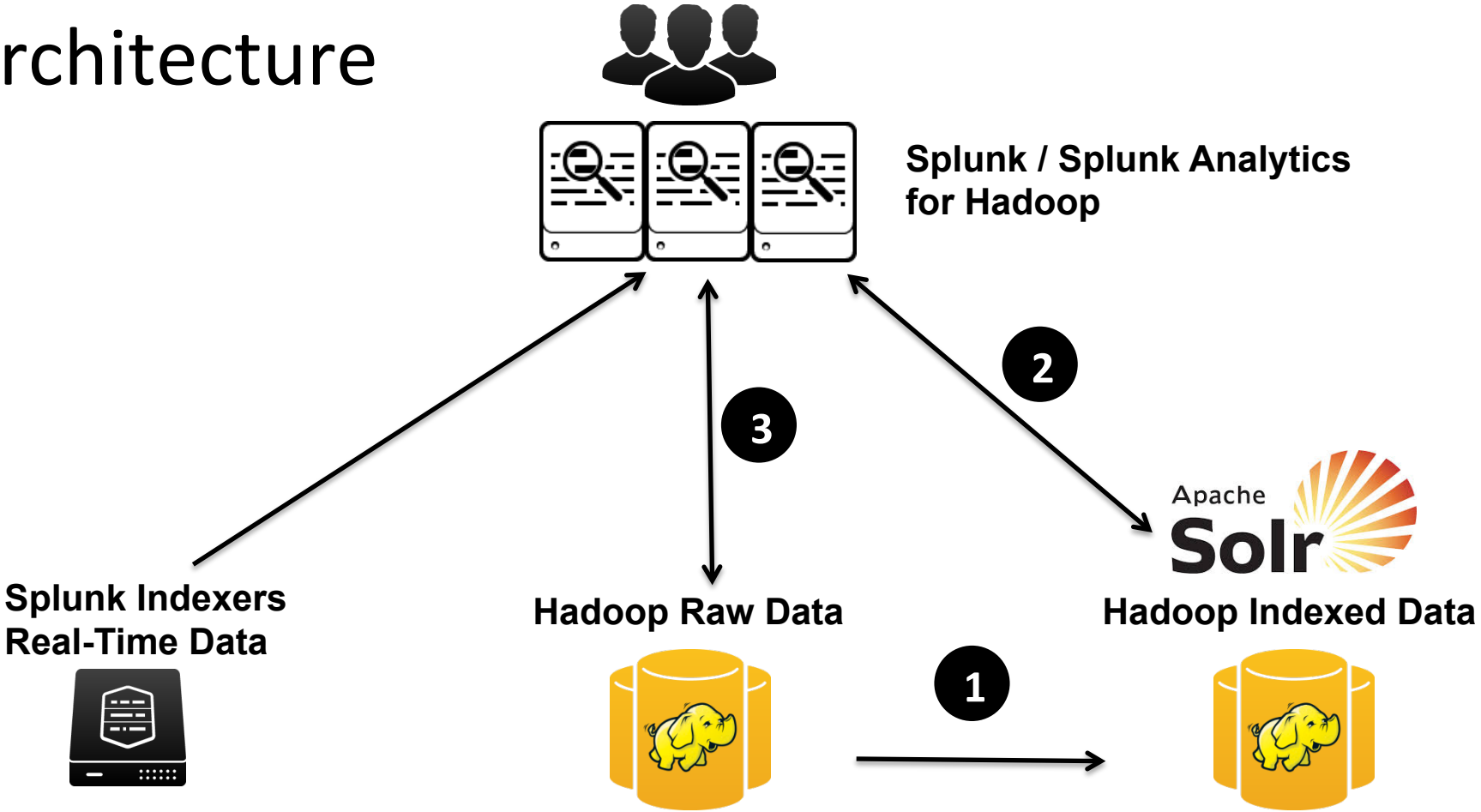
Fraud With Solr, Splunk, And Splunk Analytics For Hadoop

Use Case: Fraud – Why Apache Solr

Apache Solr is an open source enterprise search platform from **the Apache Lucene** API. Its major features include full-text search, hit highlighting, faceted search, and real-time **indexing**.

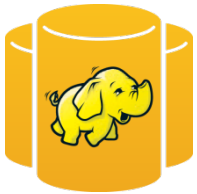
1. **Problem**: To scan a whole month of data or longer looking for a key words in Splunk Analytics for Hadoop takes long time since we have many log files to search through.
2. **Goal**: We would like to limit number of files to search and reduce number of map/reduce jobs to run.
3. **Solution**: To achieve this affect, we have created summary of key words in Solr. This Solr summary data contains key words and in which HDFS files the key words are found in.

Architecture



Fraud – Technical Details

Hadoop - Solr	Splunk – Solr	Cassandra - Splunk Analytics for Hadoop
• Solr monitors any changes to Hadoop Directory • **Index Key words based on Hadoop Source files**	• Splunk Form dashboard • User enter Key words • Python script calls Solr • **Solr returns to Splunk all Hadoop source files with the Key words**	• **Splunk Analytics for Hadoop runs MapReduce Hadoop jobs with the Specific Source Files** • Eliminates massive Hadoop scan



Hadoop Raw data

Hadoop Indexed data

Splunk Search Head

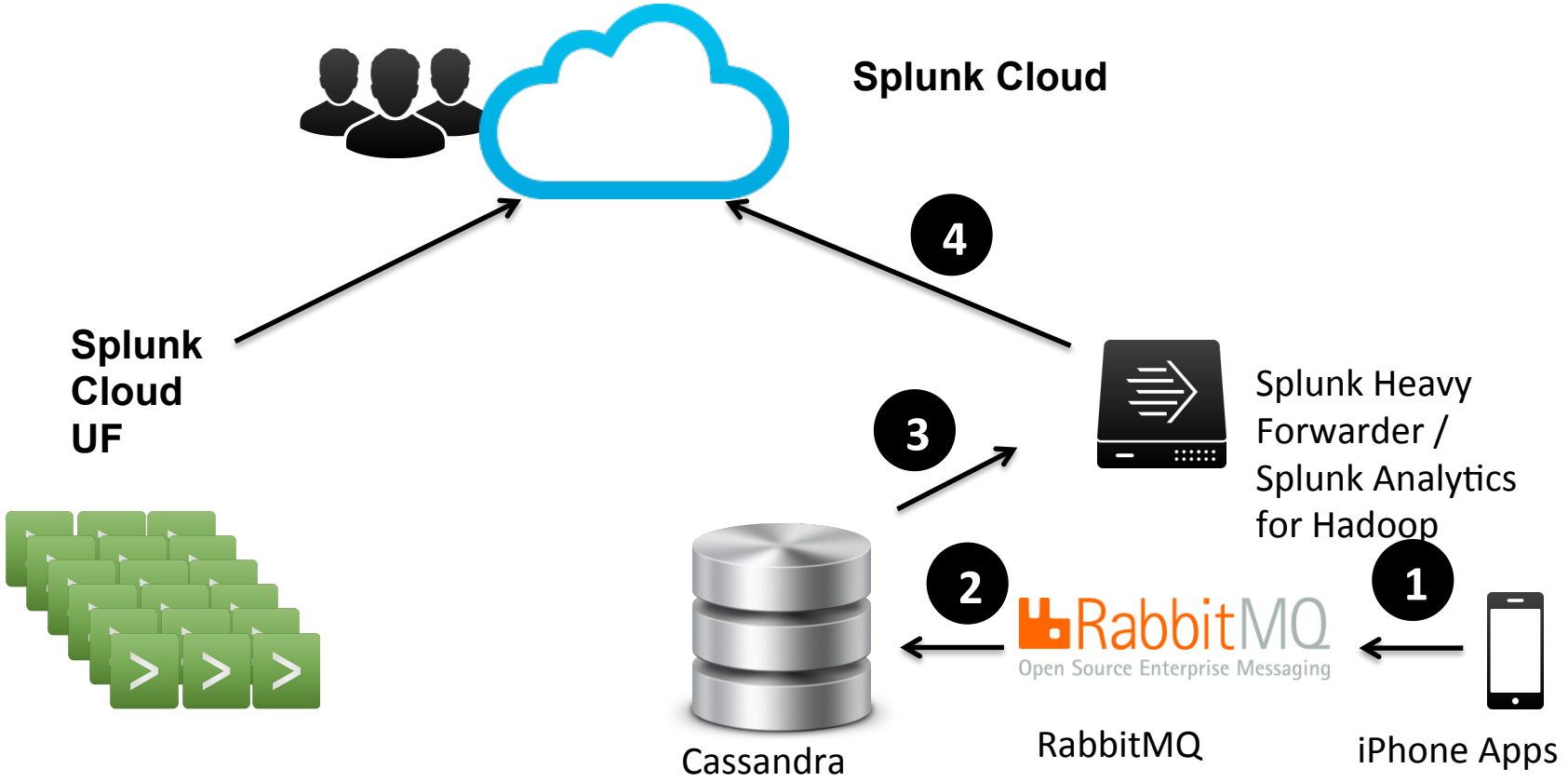
Business Analytics With Cassandra, Splunk Cloud, And Splunk Analytics For Hadoop

Business Analytics – Why Cassandra

Apache Cassandra is an open-source distributed **NoSQL database** system designed to handle large amounts of data across cluster of commodity servers.

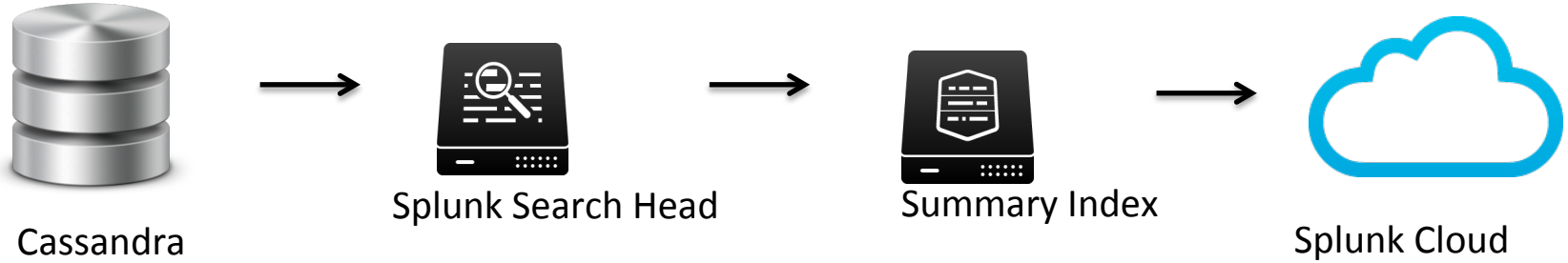
1. **Problem**: Lack of Visibility into customer behavior from Mobile Applications.
2. **Goal**: Visualize and analyze all data that is stored in Cassandra.
3. **Solution**: To achieve this affect, we stored all Mobile activity into Cassandra and use Splunk Analytics for Hadoop add on for Cassandra to query that data.

Architecture

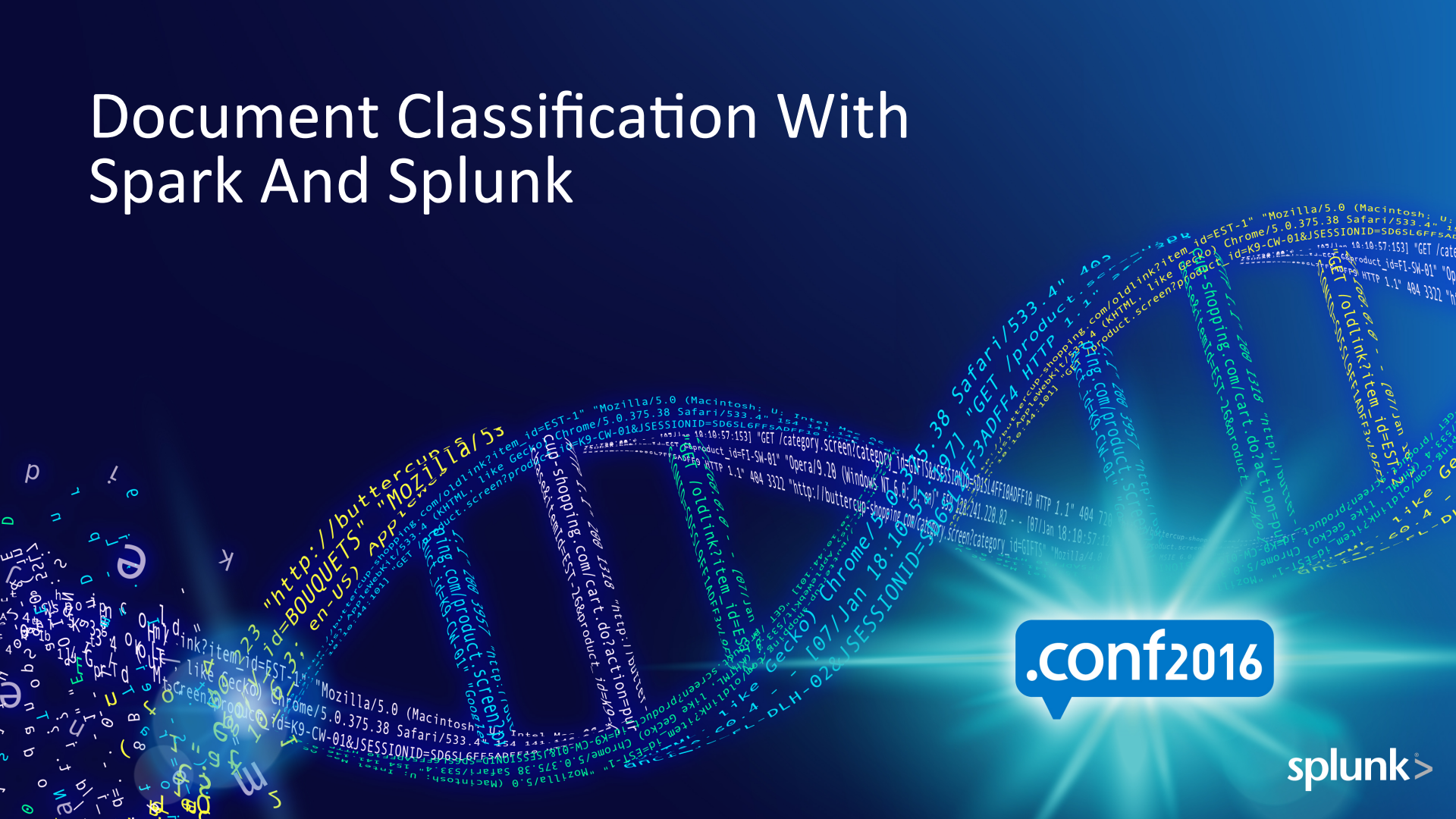


Business Analytics – Technical Details

Cassandra - Splunk Analytics for Hadoop	Splunk Analytics for Hadoop – Summary Index	Summary Index – Splunk Cloud
• **Splunk Analytics for Hadoop Add On for Cassandra** • [cassandra_weathercql] vix.provider = cassandra_erp vix.cassandra.cql.cmd = SELECT * FROM weathercql.monthly	• index = cassandra_weathercql | table * And Schedule Search • index = cassandra_weathercql | Collect **SummaryIndex**	• Output.conf [tcpout] forwardedindex.0.whitelist = SummaryIndex • SummaryIndex for 5 Min • Use the normal **Splunk Cloud UF**



Document Classification With Spark And Splunk

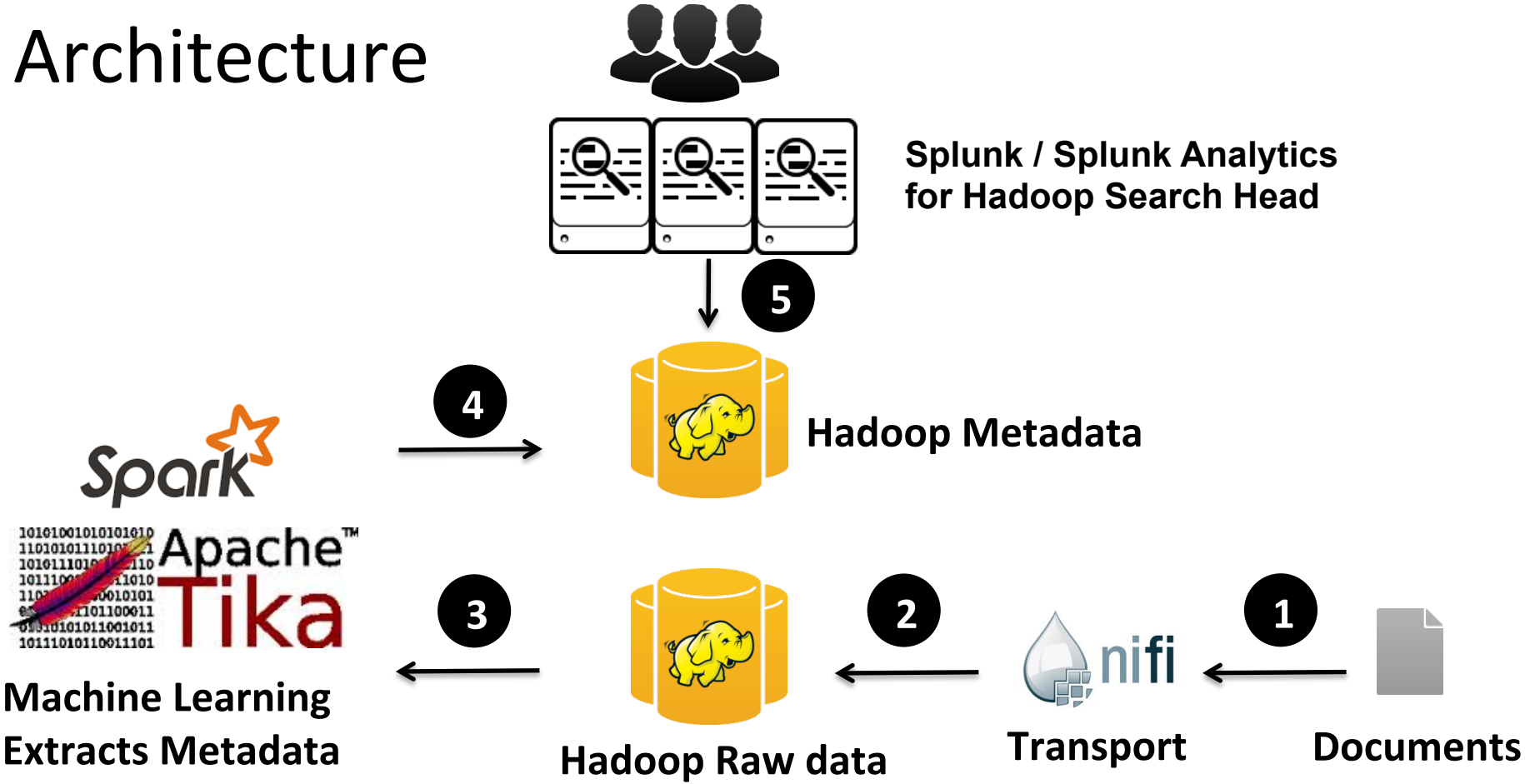


Document Classification – Why Spark

Apache Spark provides APIs that provides a fast processing and it was developed in response to limitations into the Hadoop MapReduce cluster computing paradigm. The main components for Spark are: Core, SQL, Machine Learning, Stream, and Graph APIs..

1. **Problem**: Spark processing does not provides an easy analytics or any visualization.
2. **Goal**: Allows analysts and regulators the ability to know exactly where each file exists in the system.
3. **Solution**: Apache Nifi collect all new files from NFS and store it on Hadoop. Spark Core, Spark Machine Learning, and Apache Tika creates Metadata classification. Splunk Analytics for Hadoop expose Metadata classification files to end users.

Architecture



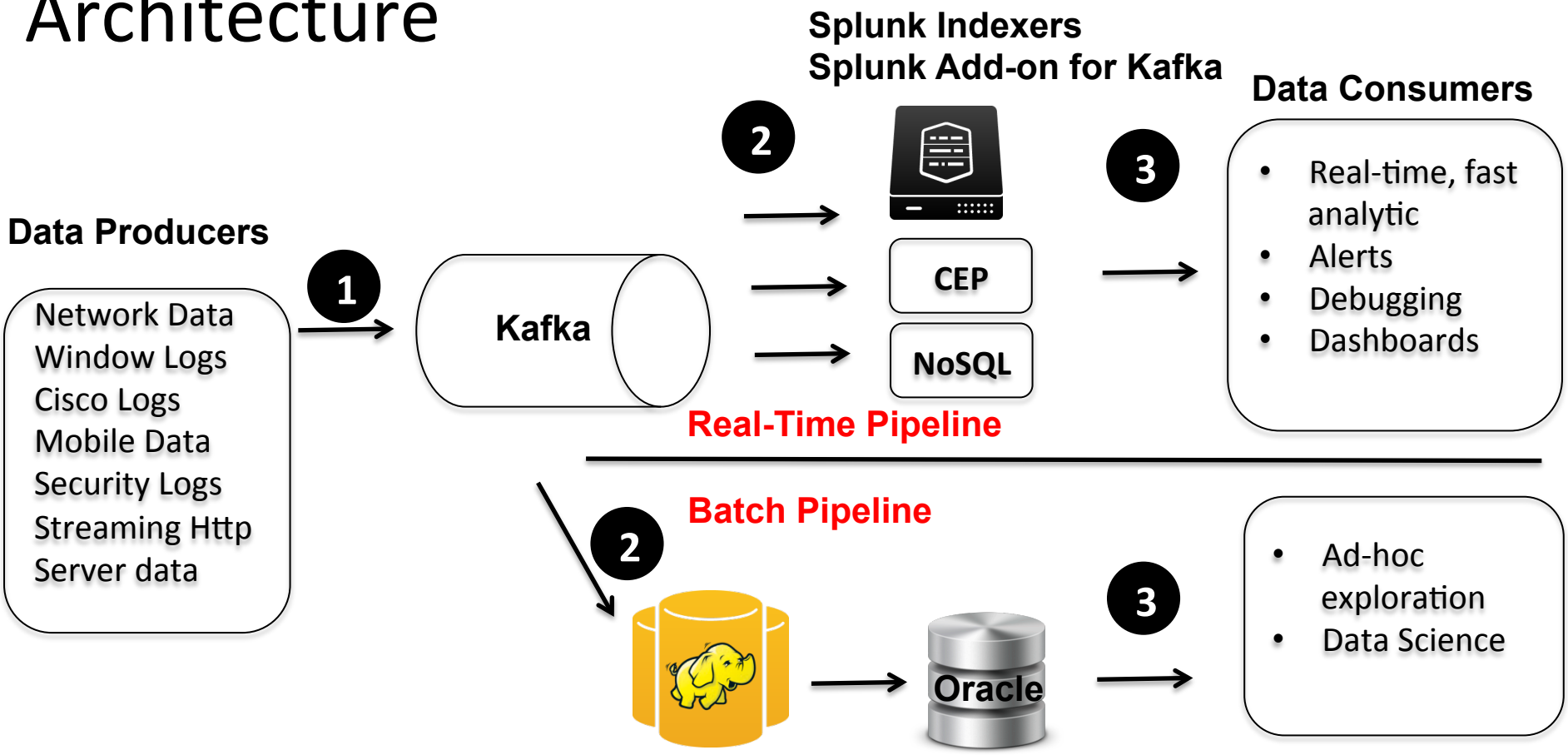
Network IT With Kafka And Splunk Kafka Add On

Network IT – Why Kafka

Apache Kafka is a fast publish-subscribe messaging system. A single Kafka broker can handle hundreds of megabytes of reads and writes per second from thousands of clients a indexing.

1. **Problem**: No unified collection framework.
2. **Goal**: Real Time visualization and analytics using Splunk, Batch visualization and analytics using Hadoop and RDBMS.
3. **Solution**: To achieve this affect, we used a Splunk Add-on for Kafka.

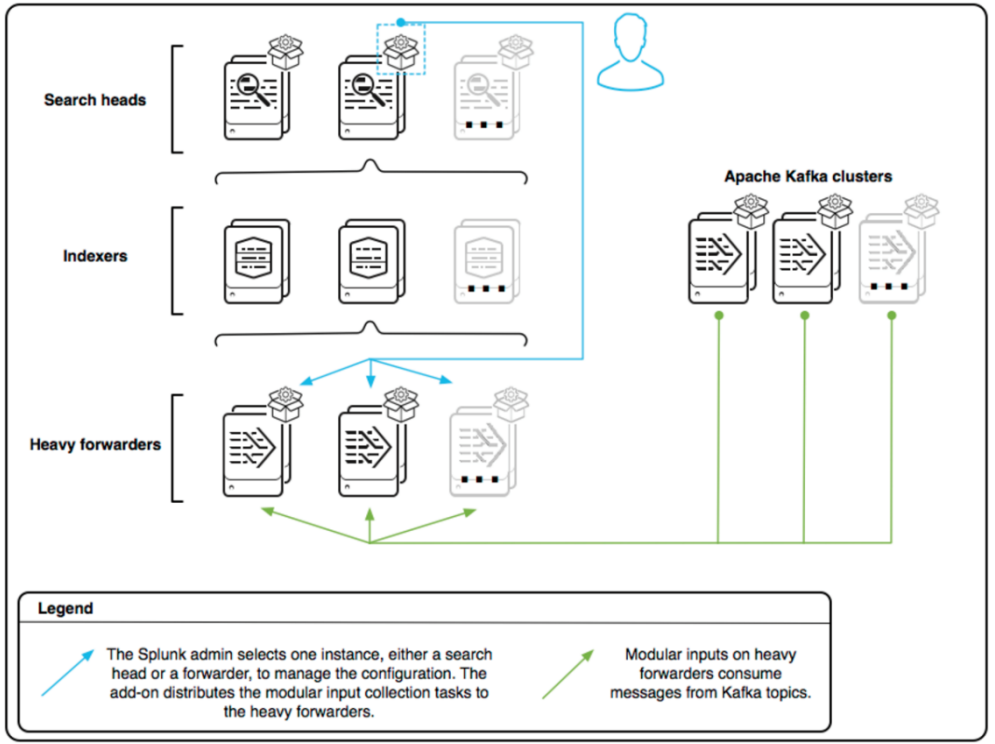
Architecture



Network IT – Technical Details

Kafka - Splunk Add-on for Kafka

- **Splunk Add-on for Kafka**
  - index = networkdata
  - kafka_brokers = sandbox:6667
  - kafka_partition_offset = earliest
  - kafka_topic_whitelist = truckevent
- **Search**
  - index=networkdata sourcetype="kafka:topicEvent"



Demo

Additional Resources

Use Cases:

- Fraud with Solr: https://lucidworks.com/solutions/lucidworks-splunk-connector/
- Business Analytics with Cassandra: https://splunkbase.splunk.com/app/2668/
- Document classification with Spark: https://splunkbase.splunk.com/app/2686/ (Spark SQL)
- Network IT with Kafka: https://splunkbase.splunk.com/app/2935/

THANK YOU

splunk>