.conf18

splunk>

Real-World Use Cases with Splunk Federated Search and Open Source Software

Raanan Dagan, Principal Sales Engineering Architect, Splunk

Bruce Penn, Senior Sales Engineer, Splunk

October 2018

Forward-Looking Statements

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 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.

Industry Experience

Raanan Dagan

- 30+ years of IT experience
- 6+ years at Splunk
- 1+ years at Cloudera (Hadoop, OSS)
- 9+ years at Oracle
- Lots of years, lots of startups

Bruce Penn

- 25+ years of IT experience
- 2+ years at Splunk
- 4+ years at MapR (Hadoop, OSS)
- 8+ years at Oracle
- Lots of years following Raanan

Agenda

Use Cases:

1. **Fraud** with Solr, Splunk, and Splunk Analytics for Hadoop
2. **Cybersecurity Posture** with Spark and Splunk Federated Search
3. **Business Analytics** with Cassandra, Splunk Cloud, Splunk Analytics for Hadoop, and Rabbit MQ
4. **Document Classification** with Apache Nifi, Spark Core, Spark Machine Learning, Apache Tika, and Splunk Analytics for Hadoop
5. **Network IT** with Kafka and Splunk Connect for Kafka

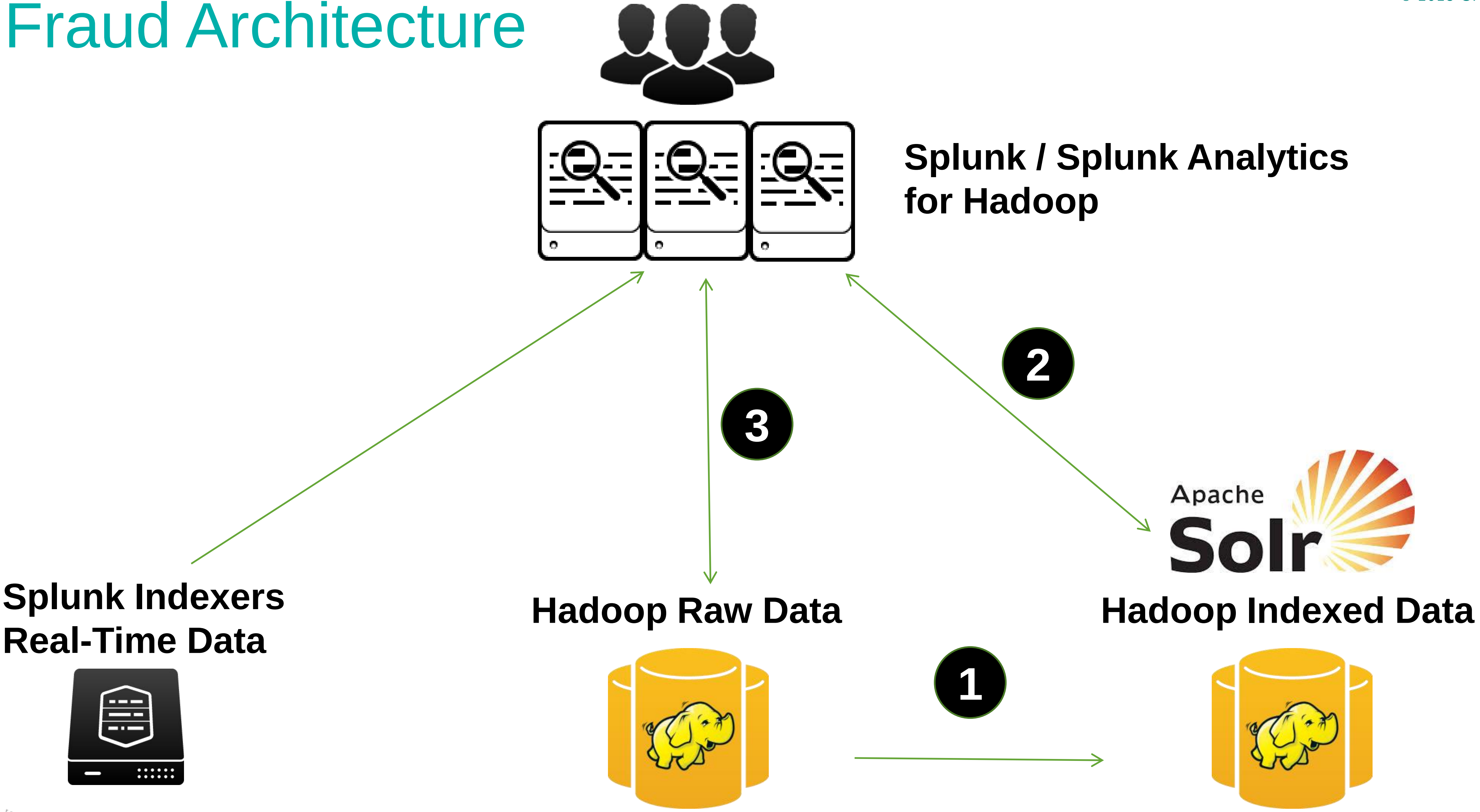
Fraud with Solr, Splunk, and Splunk Analytics for Hadoop

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**: Scanning a month or more of data while searching for keywords across Hadoop files can be a lengthy process.
2. **Goal**: Limit the number of files to search and minimize the number of MapReduce jobs to run.
3. **Solution**: Apache Solr keyword/file index and Splunk Analytics for Hadoop.

Fraud Architecture



Splunk / Splunk Analytics for Hadoop

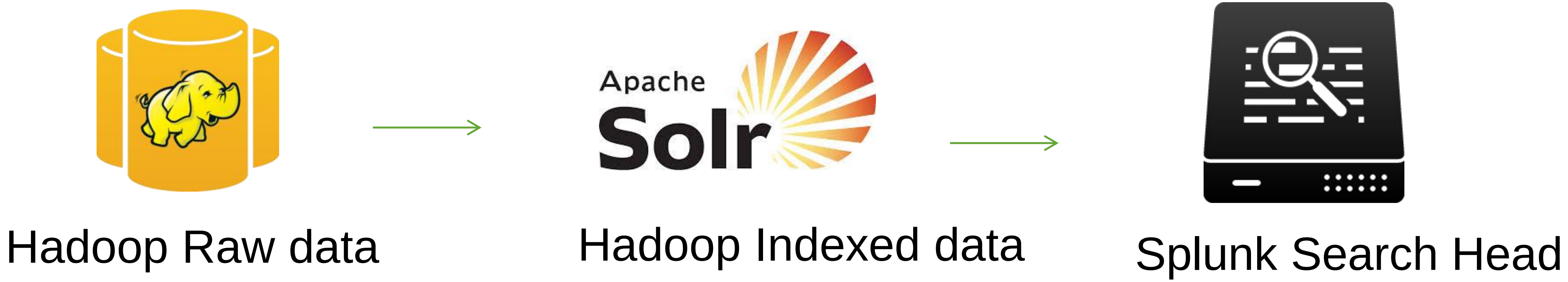
Splunk Indexers Real-Time Data

Hadoop Raw Data

Hadoop Indexed Data

Fraud – Technical Details

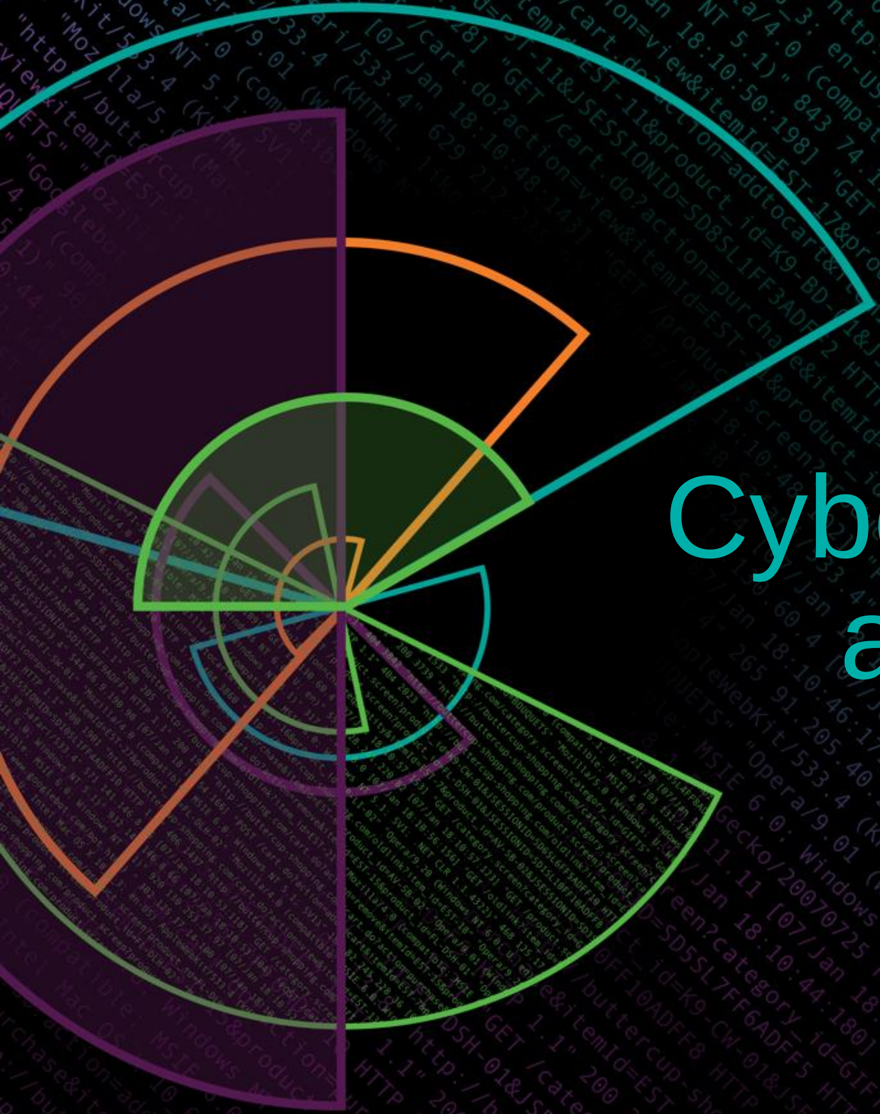
Hadoop - Solr	Splunk - Solr	Splunk Analytics for Hadoop
• Solr monitors changes to Hadoop directory/files • Indexes keywords based on Hadoop files	• Splunk form dashboard • User enters keyword(s) • Python script calls Solr • Solr tells Splunk all Hadoop files with keywords	• Splunk Analytics for Hadoop runs MR jobs with targeted files • Eliminates massive Hadoop scan



Hadoop Raw data

Hadoop Indexed data

Splunk Search Head



Cybersecurity Posture with Spark and Splunk Federated Search

Cybersecurity Posture – Why Spark?

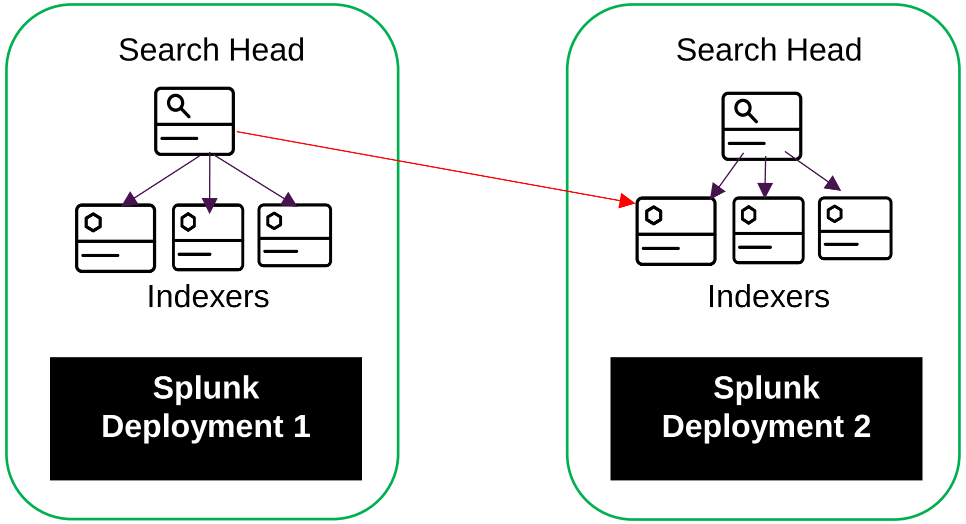
Apache Spark provides APIs that provide very fast, in-memory processing and was developed in response to limitations with the Hadoop MapReduce cluster computing paradigm.

1. **Problem**: Spark processing does not provide easy analytics or any visualization.
2. **Goal**: Allow Security Analysts access to multiple Splunk implementations while maintaining Knowledge Object and Access Control.
3. **Solution**: Spark and Splunk Federated Search.

Federated Searches across Splunk

Reaching data where it resides - other Splunk Deployments

- **Before Federated Search with multiple deployments:** Unified Search across Splunk Deployments



Search Head

Indexers

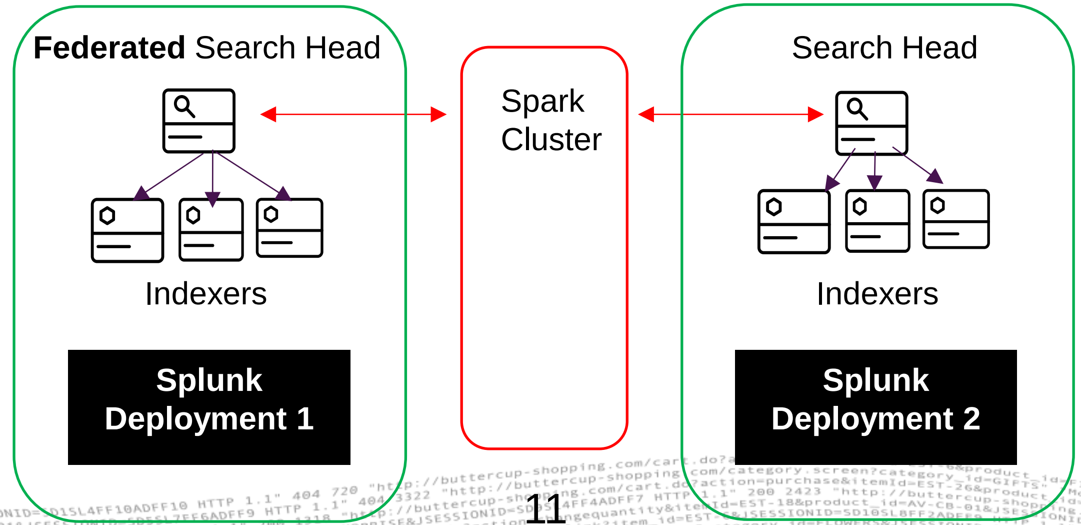
Splunk Deployment 1

Search Head

Indexers

Splunk Deployment 2

- **After Federated Search with multiple deployments:** Seamlessly search across all deployments



Federated Search Head

Indexers

Splunk Deployment 1

Spark Cluster

Search Head

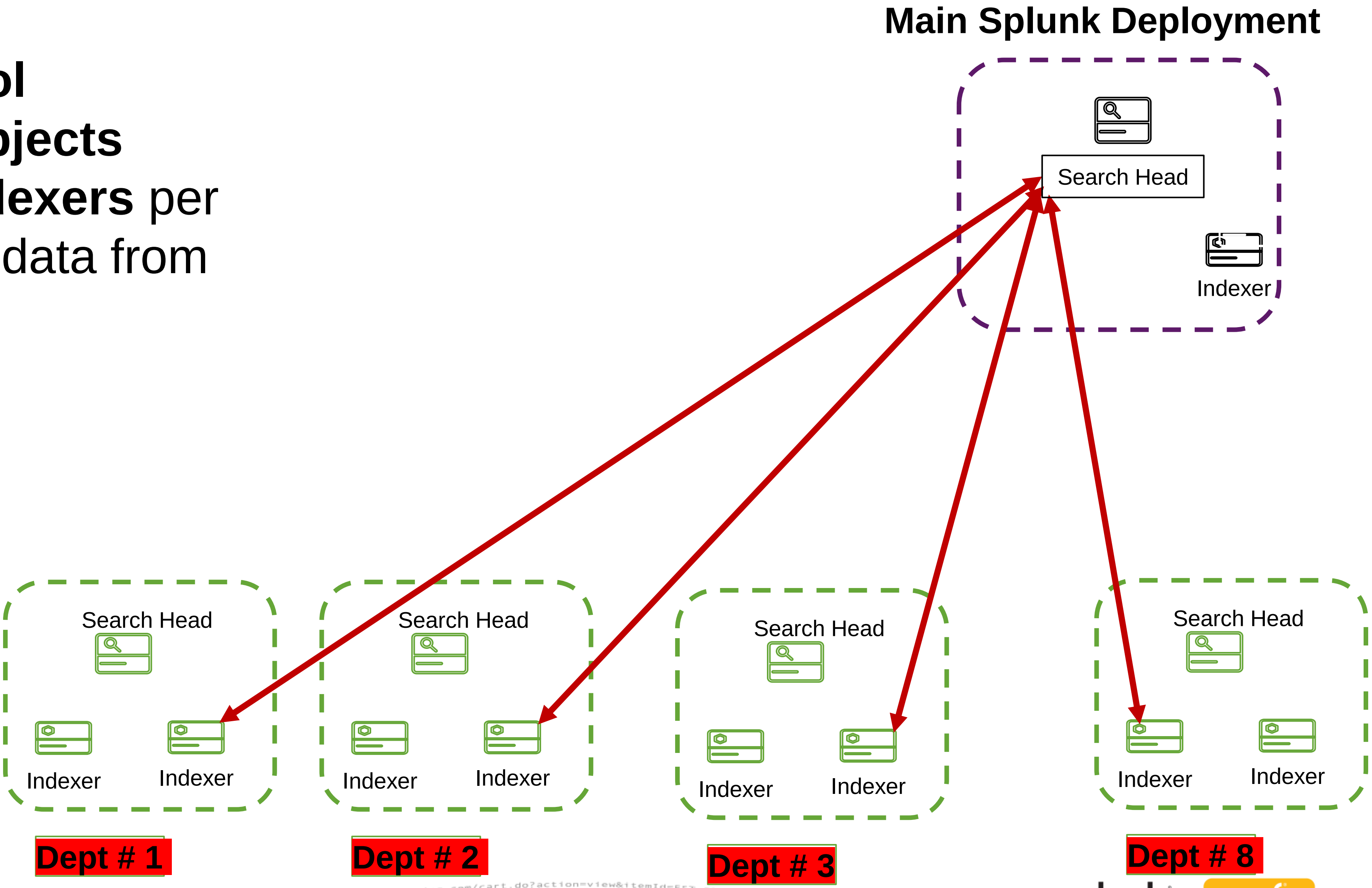
Indexers

Splunk Deployment 2

Before Federated Search

Limitations:

- **No Access Control**
- **No Knowledge Objects**
- **Need for extra indexers** per deployment (Copy data from original indexers)

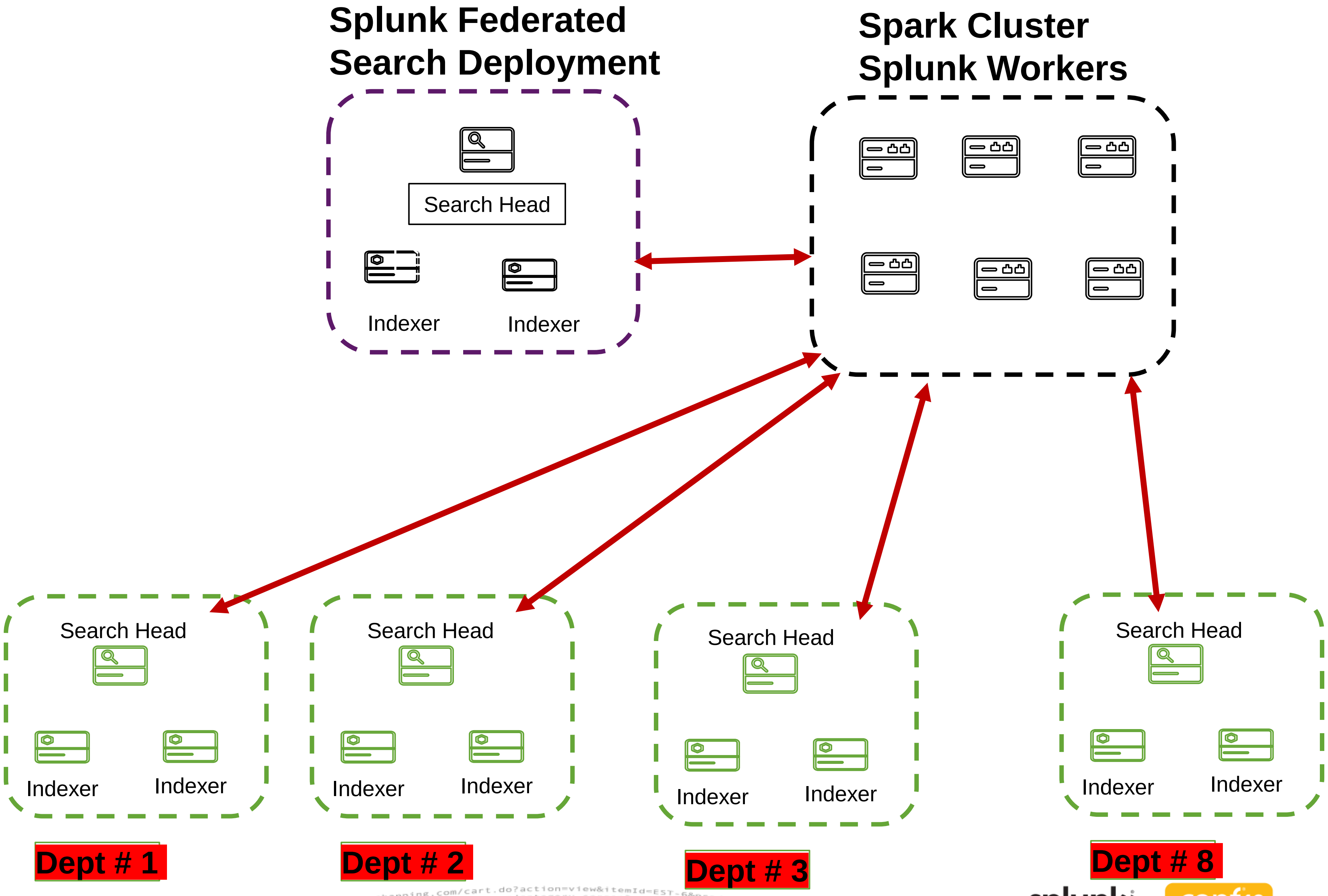


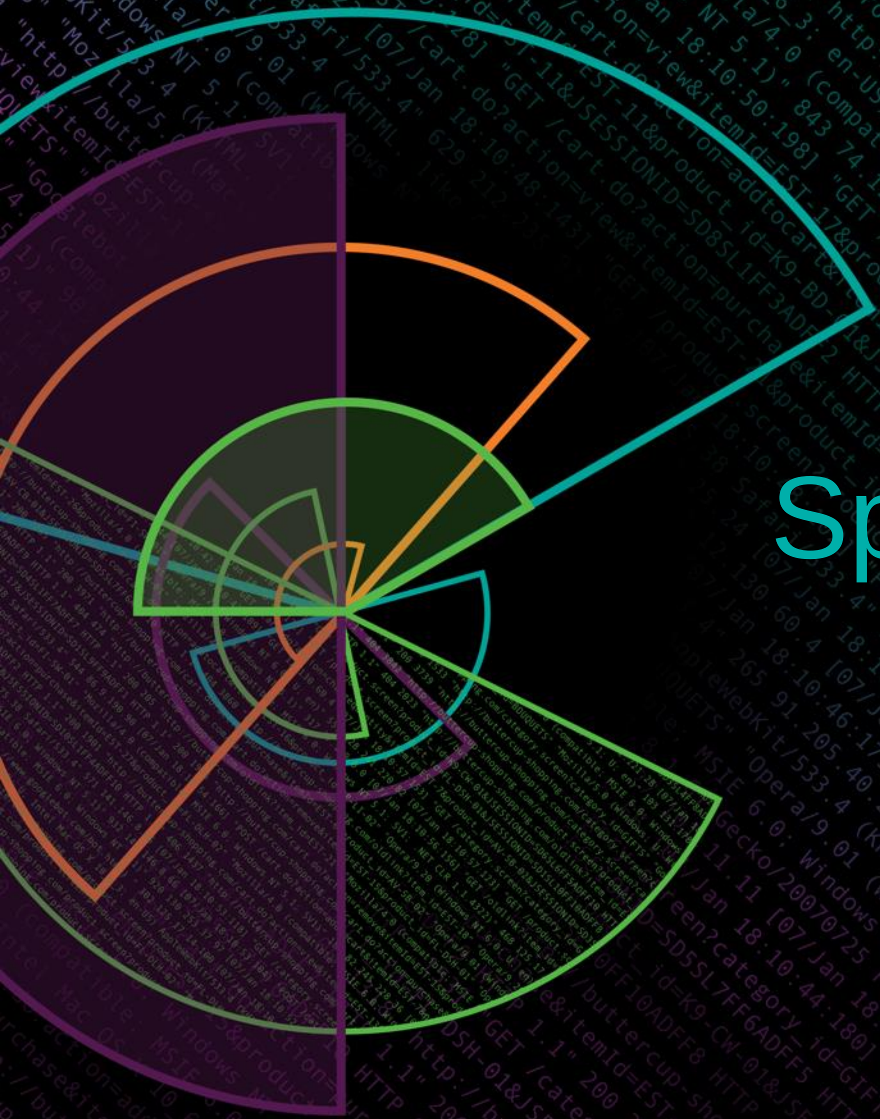
After Federated Search with Spark

Spark and Federated Search Architecture:

- Access Control
- Knowledge Objects
- Network Security
- Massive Scale compute
- No need for extra indexers per deployment

- Example Splunk search |union [|from federated:**my_dep_1_search_1**] [|from federated:**my_dep_2_search_2**] | stats count





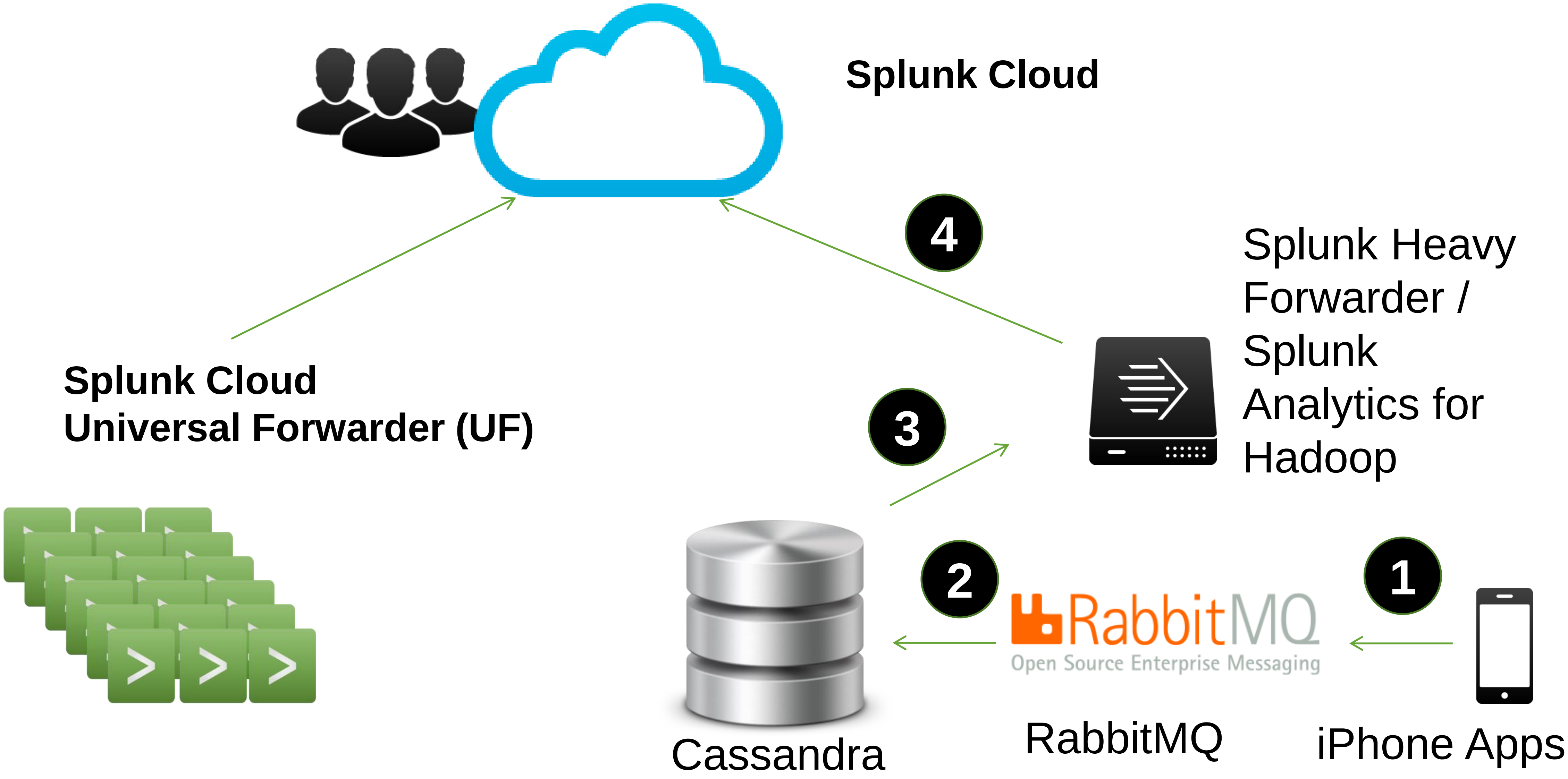
Business Analytics with Cassandra, Splunk Cloud, Splunk Analytics for Hadoop, and Rabbit MQ

Business Analytics – Why Cassandra?

Apache Cassandra is an open-source distributed **NoSQL database** system designed to handle large amounts of data across a 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**: Store all mobile activity into Cassandra via Rabbit MQ and use Splunk Analytics for Hadoop with Cassandra External Result Provider (ERP) to query that data.

Business Analytics Architecture



Splunk Cloud

Splunk Cloud
Universal Forwarder (UF)

4

Splunk Heavy Forwarder / Splunk Analytics for Hadoop

3

Cassandra

2

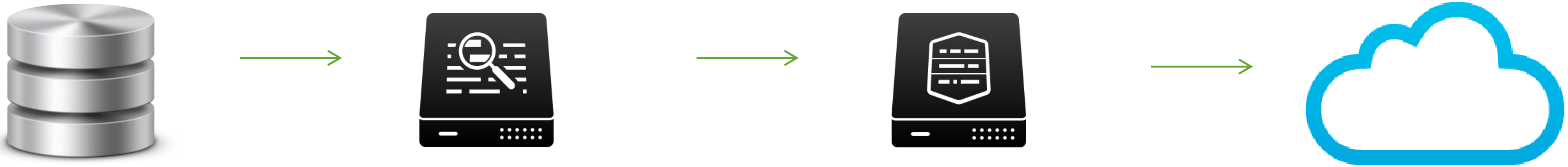
RabbitMQ

1

iPhone Apps

Business Analytics – Technical Details

Cassandra - Splunk Analytics for Hadoop	**Splunk Analytics for Hadoop – Summary Index**	**Summary Index – Splunk Cloud**
• Splunk Analytics for Hadoop with Cassandra ERP [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



Cassandra → Splunk Search Head → Summary Index → Splunk Cloud

splunk> .conf18

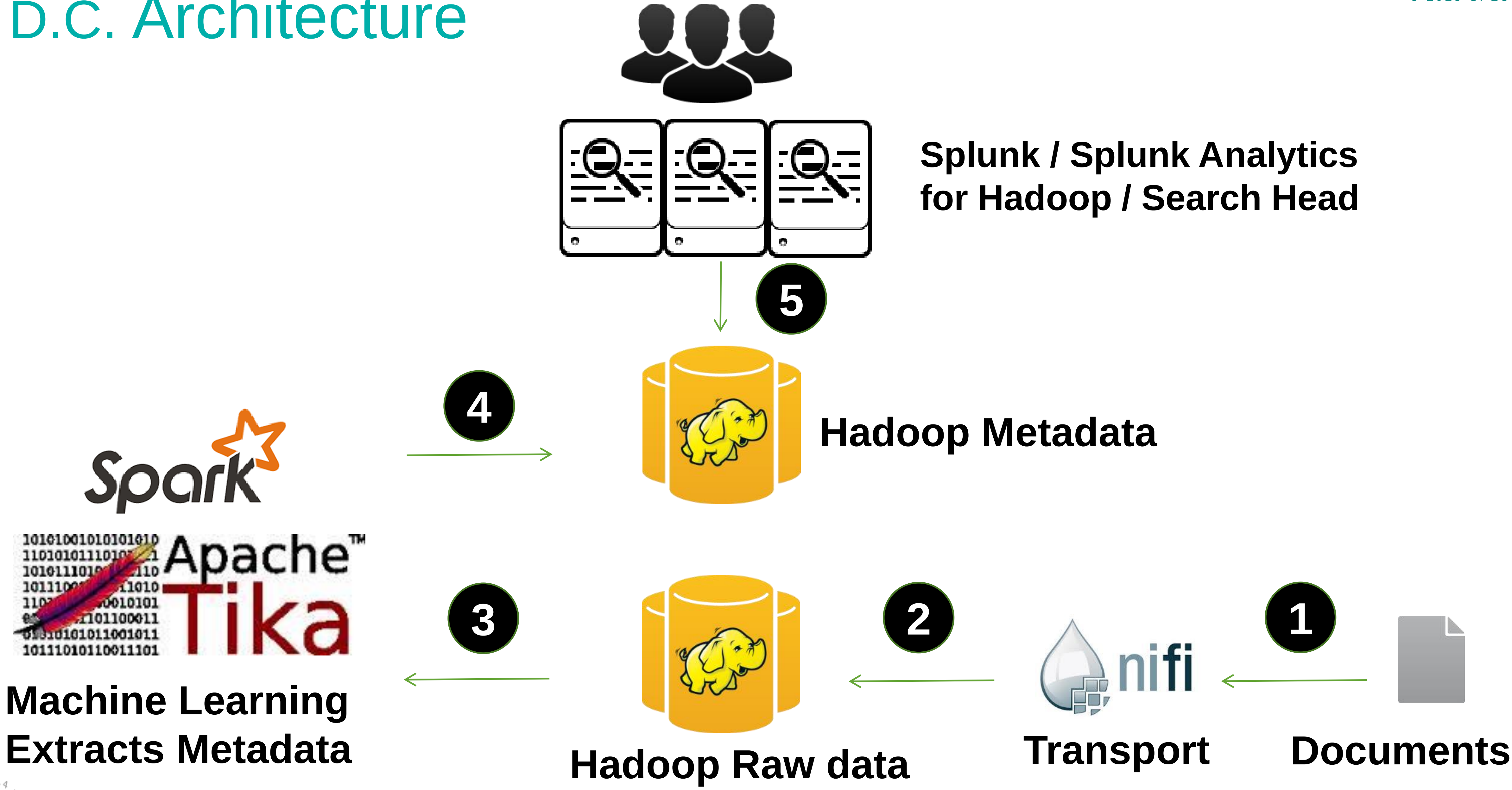
Document Classification with Apache Nifi, Spark Core, Spark Machine Learning, Apache Tika, and Splunk Analytics for Hadoop

Document Classification – Why Spark?

Apache Spark provides APIs that provide very fast, in-memory processing and was developed in response to limitations with the Hadoop MapReduce cluster computing paradigm. The main components of Spark are: Core, Spark-SQL, Machine Learning, Stream, and Graph APIs.

1. **Problem**: Spark processing does not provides 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, Spark Core, Spark Machine Learning, Apache Tika, and Splunk Analytics for Hadoop.

D.C. Architecture



Splunk / Splunk Analytics for Hadoop / Search Head

5

Hadoop Metadata

4

Machine Learning Extracts Metadata

3

Hadoop Raw data

2

Transport

1

Documents

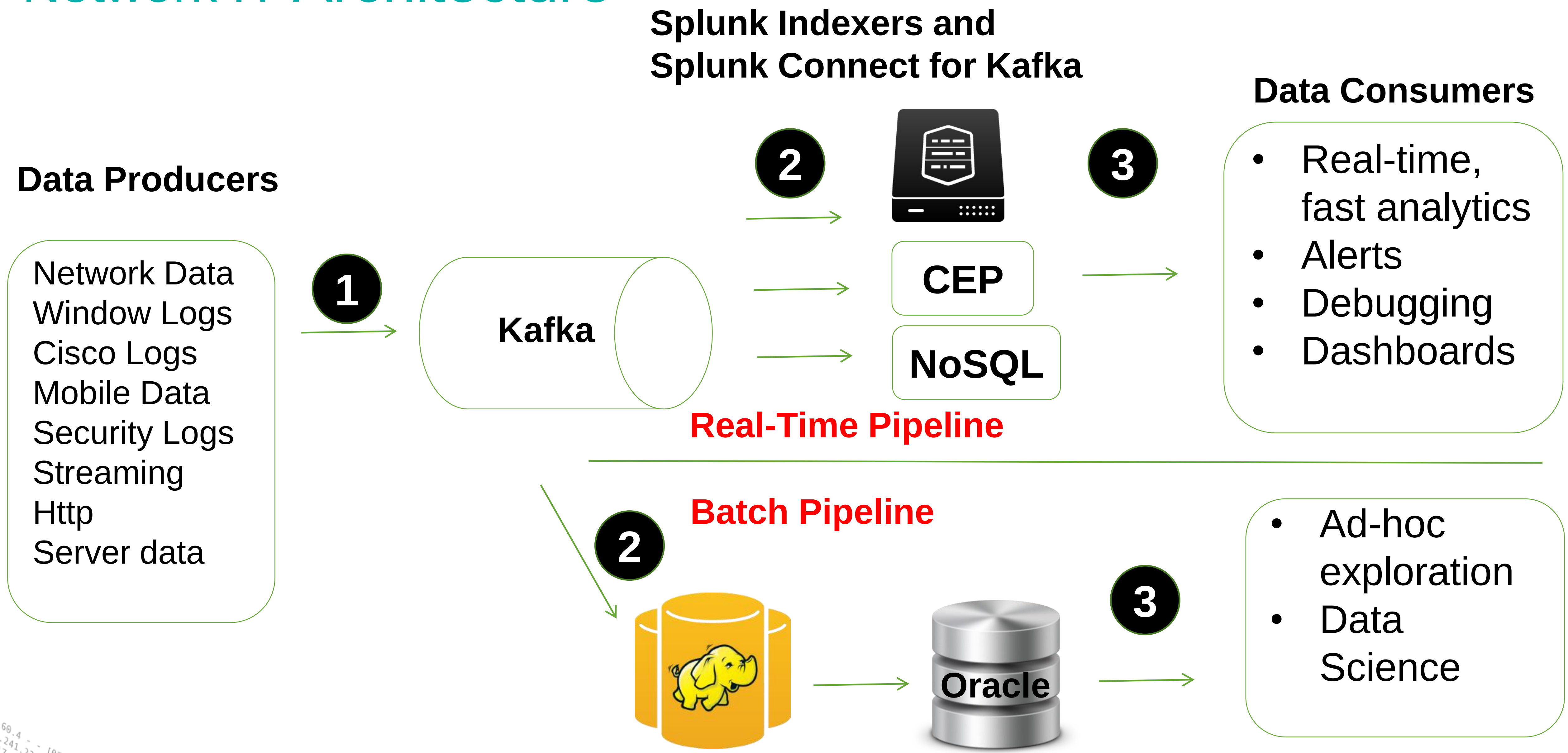
Network IT with Kafka and Splunk Connect for Kafka

Network IT – Why Kafka?

Apache Kafka is a very fast and distributed publish-subscribe messaging system. A single Kafka broker can handle hundreds of megabytes of reads and writes per second from thousands of clients while 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**: Kafka, Hadoop, Splunk Connect for Kafka, Oracle.

Network IT Architecture



Data Producers

- Network Data
- Window Logs
- Cisco Logs
- Mobile Data
- Security Logs
- Streaming
- Http
- Server data

1

Kafka

Splunk Indexers and Splunk Connect for Kafka

2

CEP

NoSQL

3

Data Consumers

- Real-time, fast analytics
- Alerts
- Debugging
- Dashboards

Real-Time Pipeline

Batch Pipeline

2

Oracle

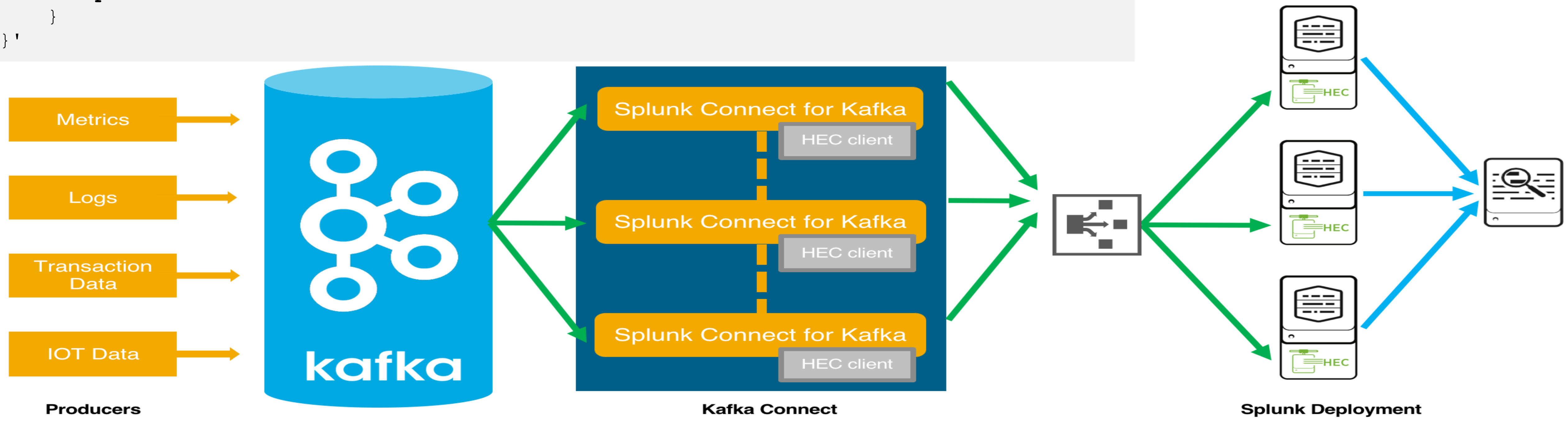
3

- Ad-hoc exploration
- Data Science

Network IT – Technical Details

```
curl <KAFKA_CONNECT_HOST>:8083/connectors -X POST -H "Content-Type: application/json" -d'{
  "name": "splunk-prod-financial",
    "config": {
      "connector.class": "com.splunk.kafka.connect.SplunkSinkConnector",
      "tasks.max": "10",
      "topics": "t1",
      "splunk.hec.uri": "https://elb-kafka:8088",
      "splunk.hec.token": "1B901D2B-576D-40CD-AF1E-98141B499534",
      "splunk.hec.ack.enabled : "true",
      "splunk.hec.raw" : "true",
      "splunk.hec.raw.line.breaker" : "#####”
      "splunk.hec.total.channels": "4"
    }
}'
```

~30 TB/day



Metrics

Logs

Transaction Data

IOT Data

Producers

Splunk Connect for Kafka

HEC client

Splunk Connect for Kafka

HEC client

Splunk Connect for Kafka

HEC client

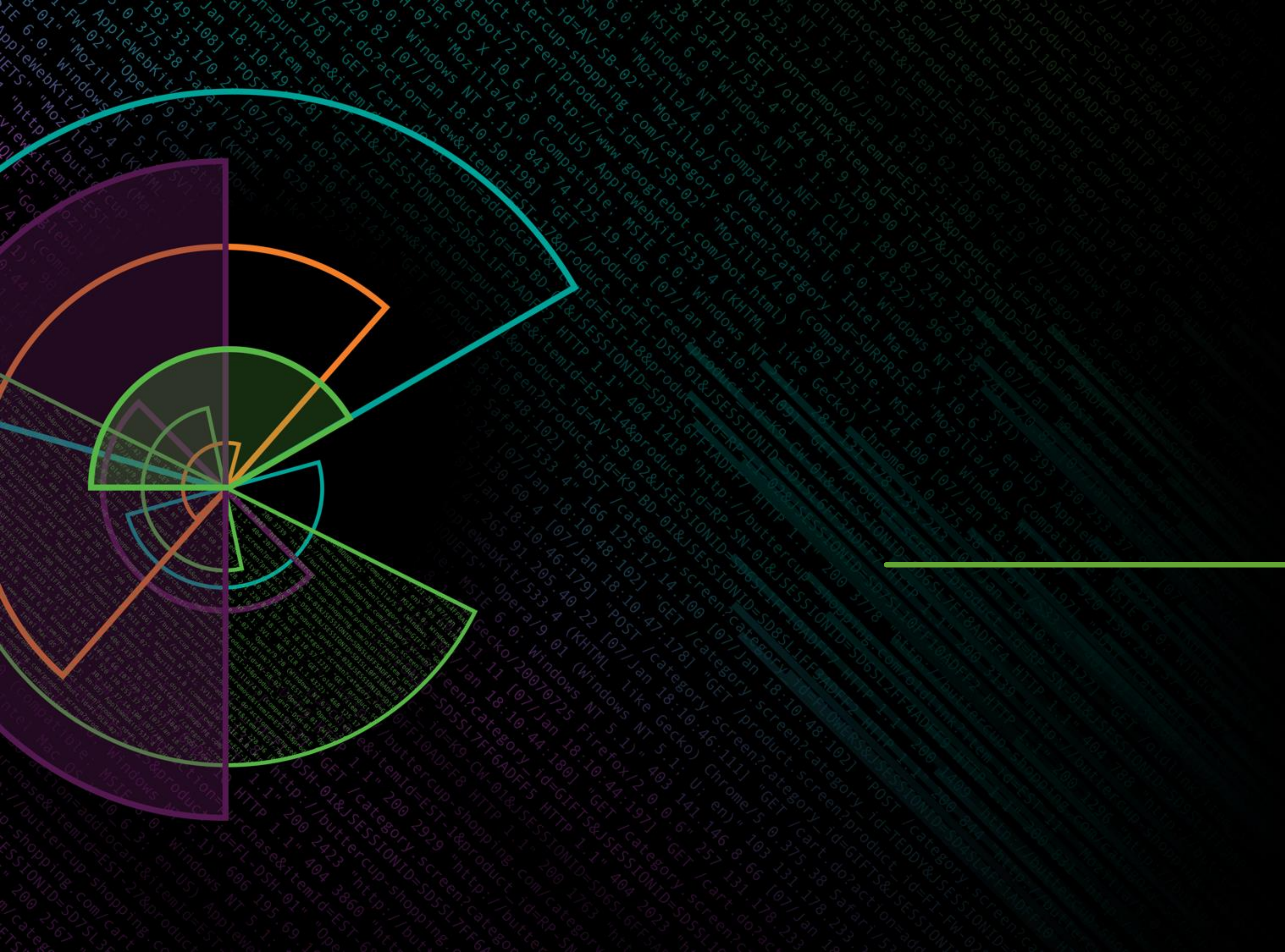
Kafka Connect

Splunk Deployment

Additional Resources

Use Cases:

1. Fraud with Solr: https://lucidworks.com/resources/#all/splunk
2. Business Analytics with Cassandra: https://splunkbase.splunk.com/app/2668/
3. Document Classification with Spark: https://splunkbase.splunk.com/app/2686/ (Spark SQL)
4. Network IT with Kafka: https://splunkbase.splunk.com/app/3862/



Q&A

Thank You!

Don't forget to rate this session in the .conf18 mobile app

.conf18

splunk>