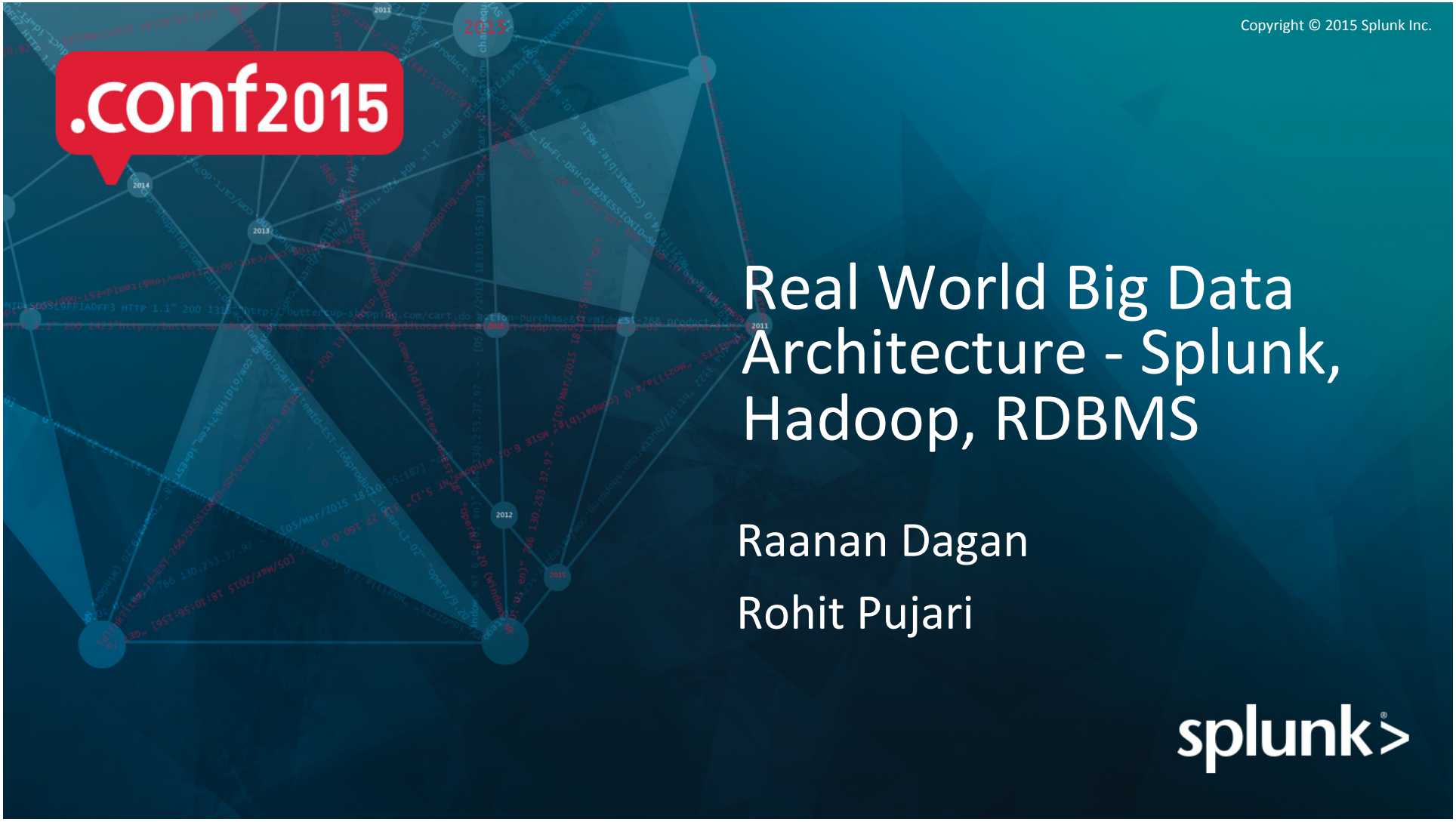


Real World Big Data Architecture - Splunk, Hadoop, RDBMS

Raanan Dagan

Rohit Pujari

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

- Splunk Big Data Architecture
- Alternative Open Source Approach
- Real-World Customer Architecture
- End-to-end Demonstration

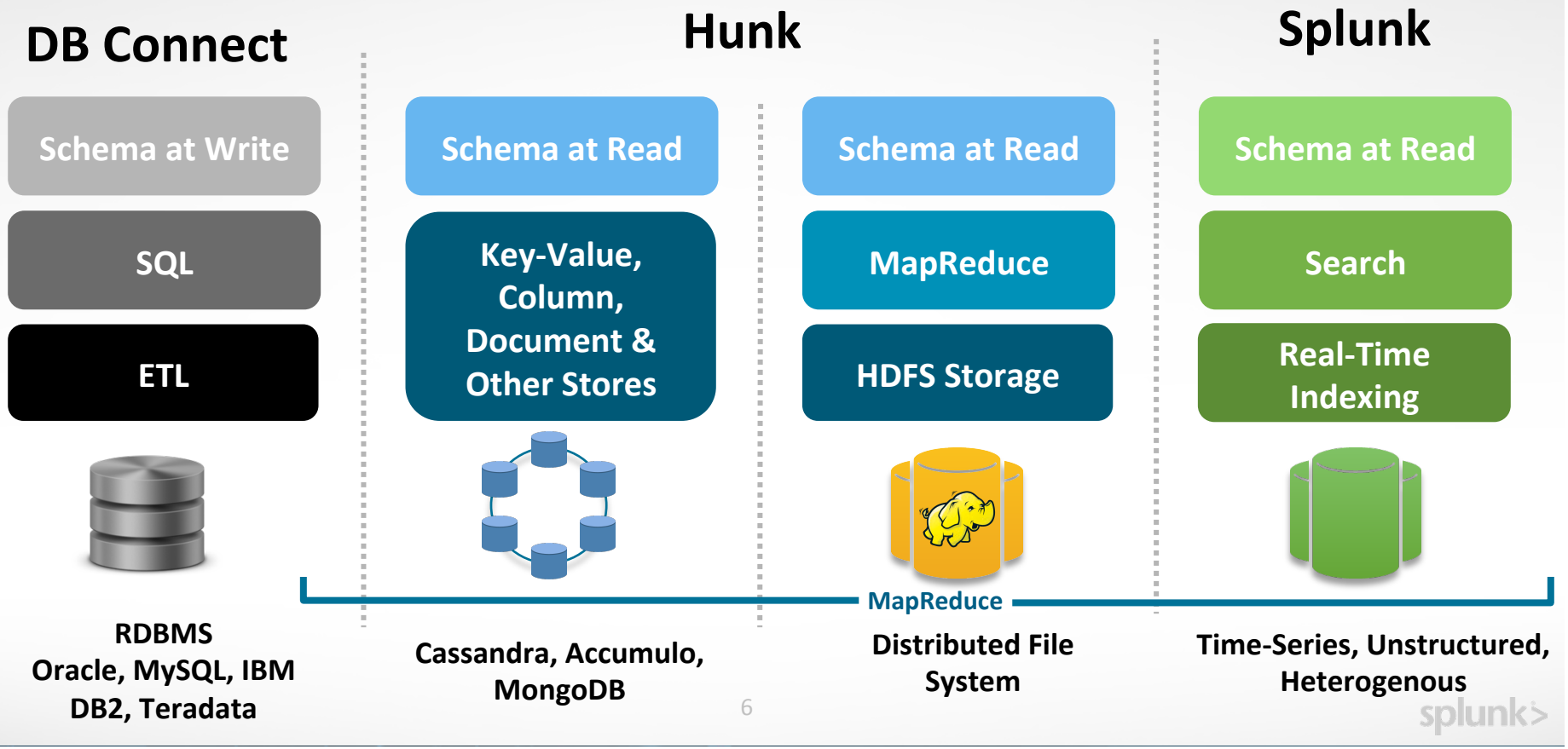
Who are you?

- Raanan Dagan - Sr. SE, Big Data specialist
- Rohit Pujari – Sr. SE, Big Data SME

Big Data Technologies

Relational Database Structured	NoSQL Semi-Structured	Hadoop Semi-Structured	Splunk
Schema at Write	Schema at Read	Schema at Read	Schema at Read
SQL	Key-Value, Column, Document & Other Stores	MapReduce	Search
ETL		HDFS Storage	Real-Time Indexing
RDBMS Oracle, MySQL, IBM DB2, Teradata	Cassandra, Accumulo, MongoDB	MapReduce Distributed File System	Time-Series, Unstructured, Heterogenous

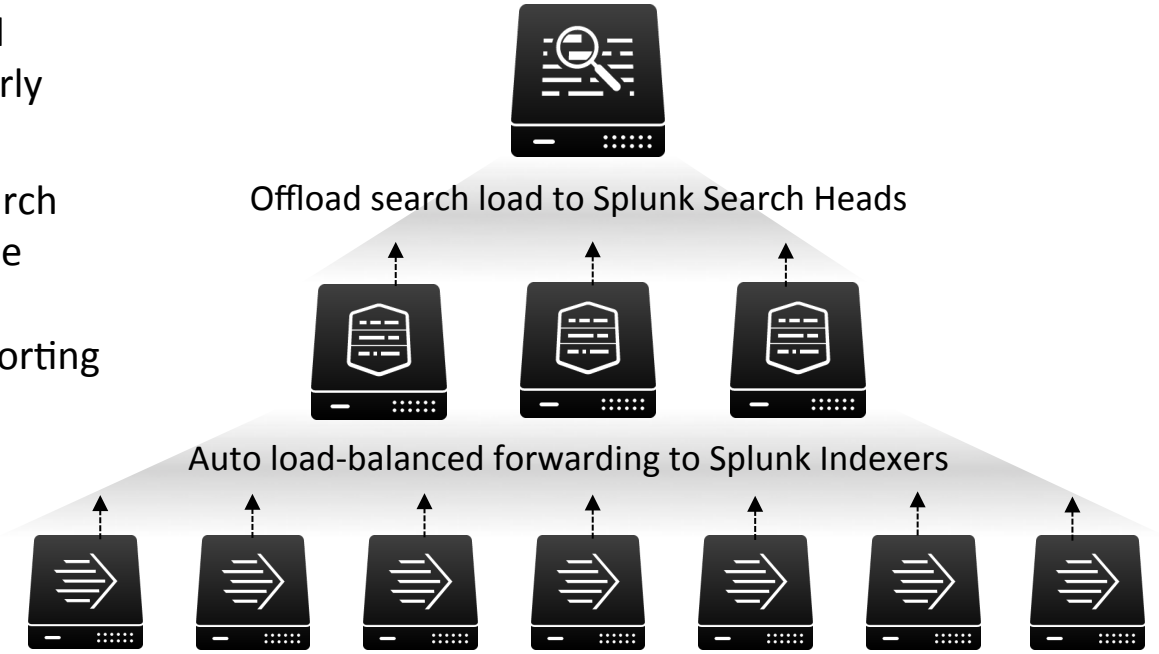
Splunk Big Data Technologies



Splunk Scalability

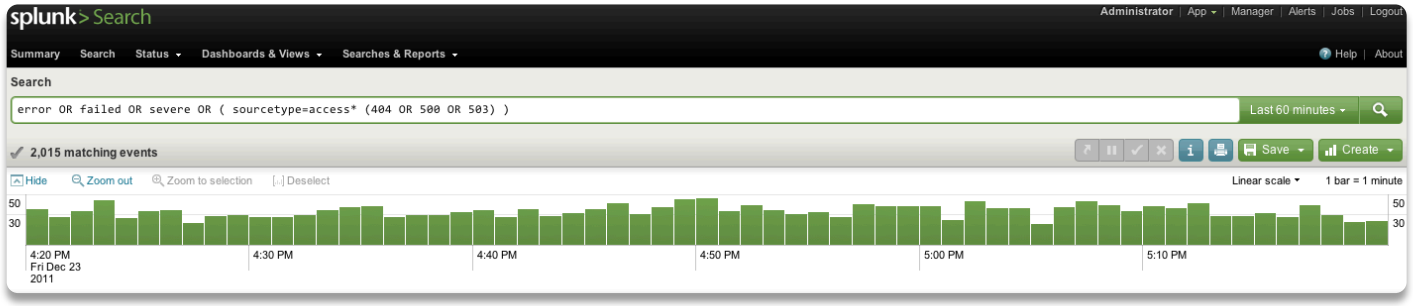
Enterprise-class Availability and Scale

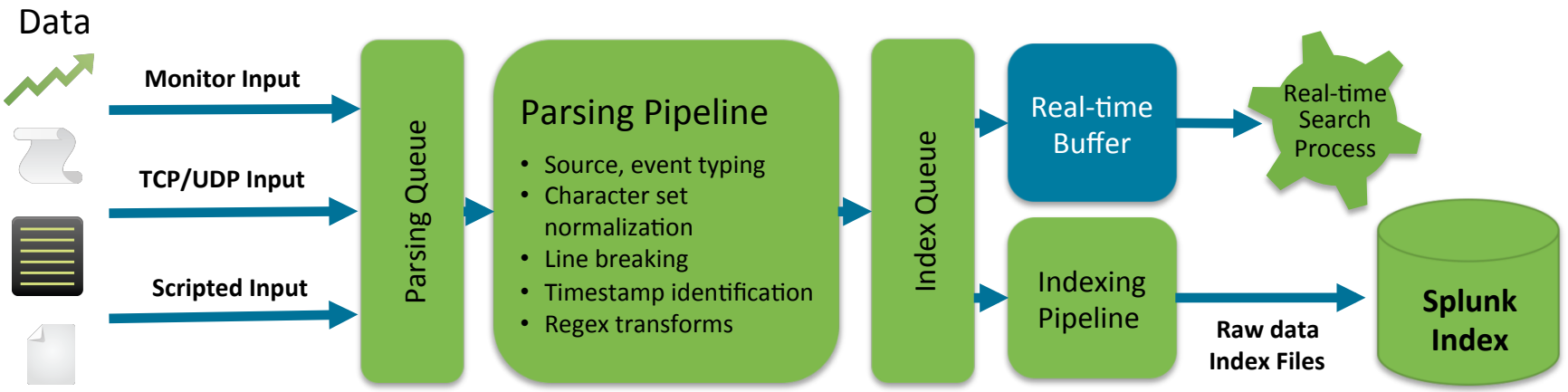
- Automatic load balancing linearly scales indexing
- Distributed search and MapReduce linearly scales search and reporting



Send data from thousands of servers using any combination of Splunk forwarders

Splunk Real-Time Analytics



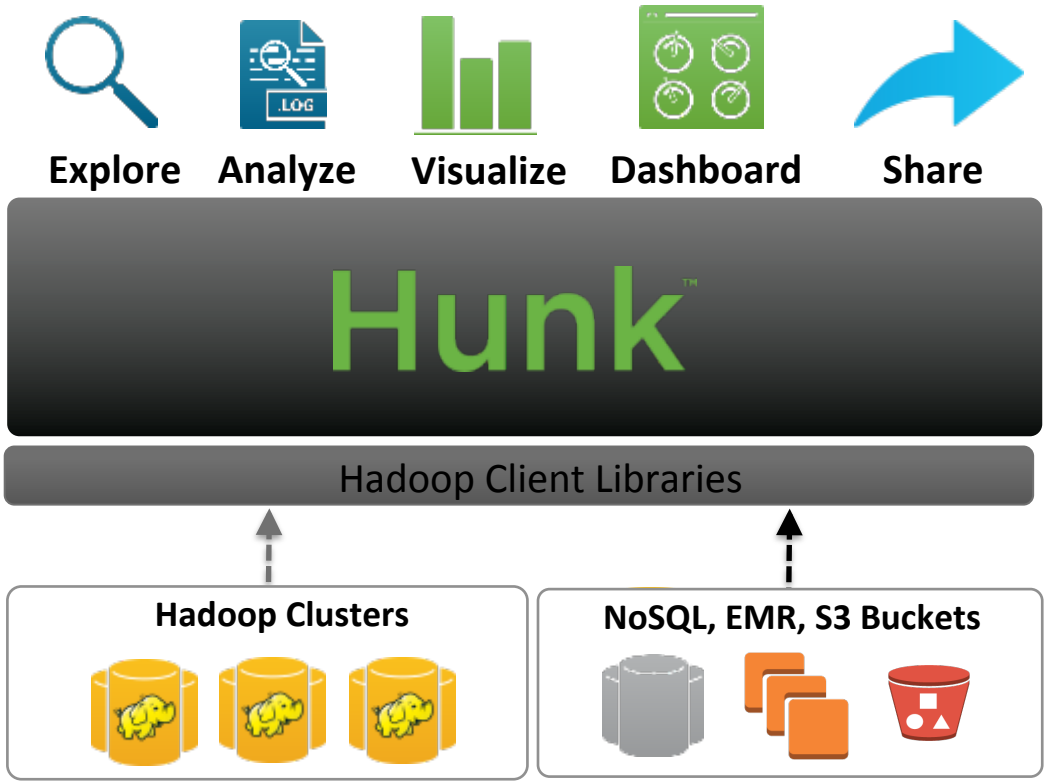


Hunk - Analytics Platform for Hadoop

Full-featured, Integrated Product

Insights for Everyone

Works with What You Have Today



Hunk Unique Features

Virtual Index

- Enables seamless use of the Splunk technology stack on data wherever it rests
- Natively handles MapReduce

Schema-on-the-fly

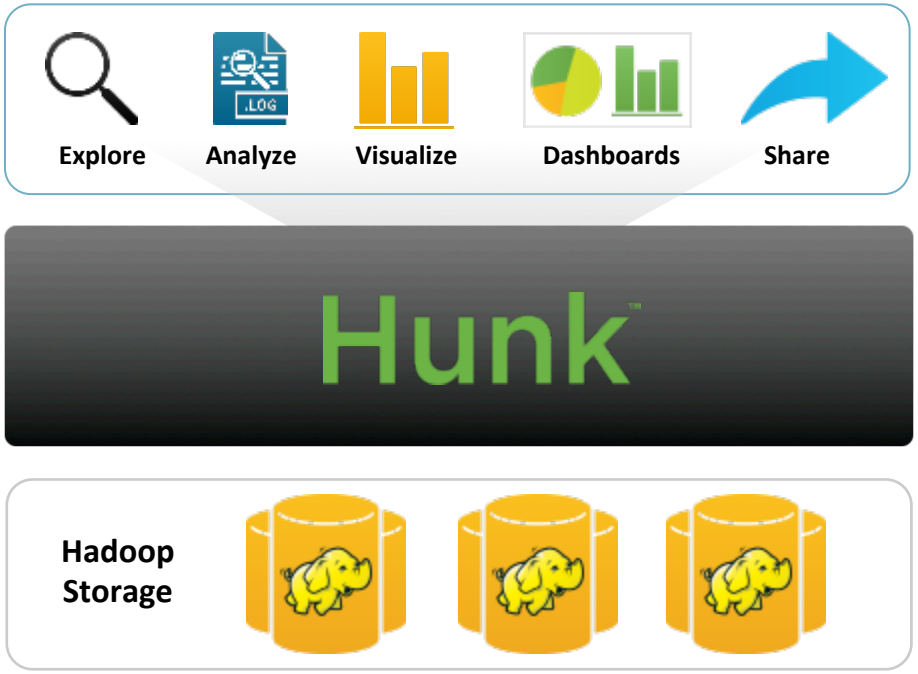
- Structure applied at search time
- No brittle schema
- Automatically find patterns and trends

Flexibility and Fast Time to Value

- Interactive search
- Preview results while MapReduce jobs run
- Drag-and-drop analytics

Security: Access Control, Pass Through Authentication, Kerberos

Hunk Provides Self-Service Analytics for Hadoop



Hunk = Indexing + Data Preview + Security + Great UI

What About Structured Data?

Customer profile	Product attributes	Employee details	Pricing and Rate plans	Asset info

Use cases for structured data in Splunk

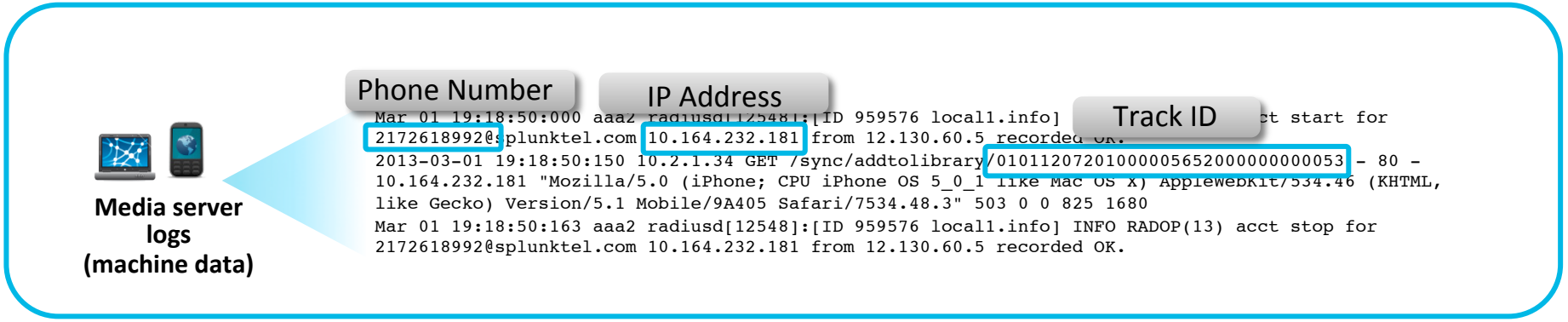
Index machine data from databases, such as logs or sales records

Enrich machine data with high-level data, such as customer records

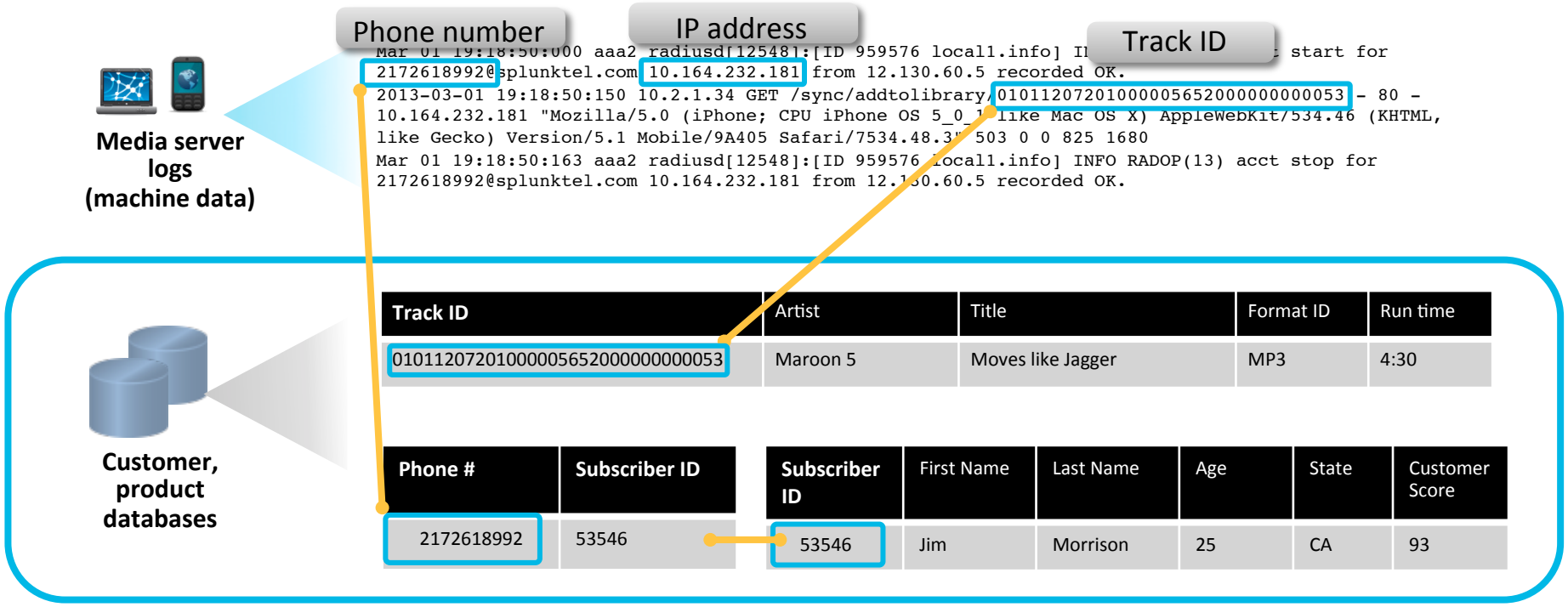
Update structured databases with Splunk info, such as risk scores

Interactively browse structured and unstructured data from Splunk reports

Machine Data – Delivers Real-time Insights



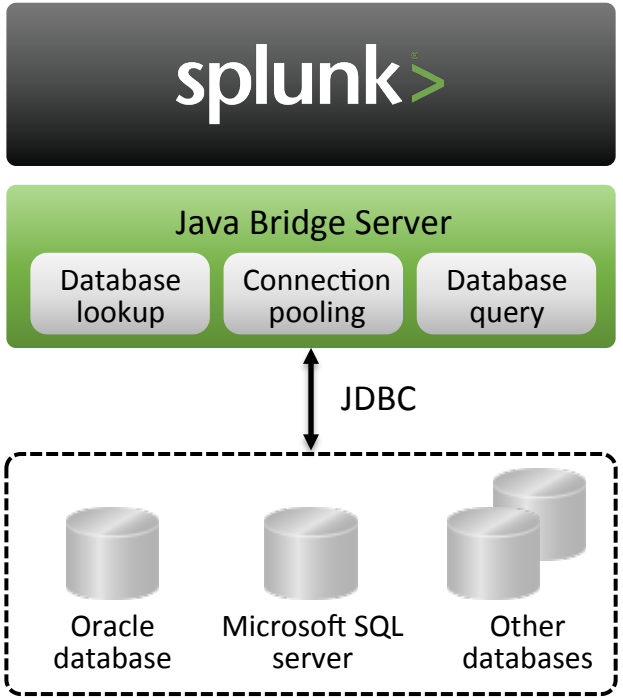
Structured Data – Contains Business Context



Splunk DB Connect

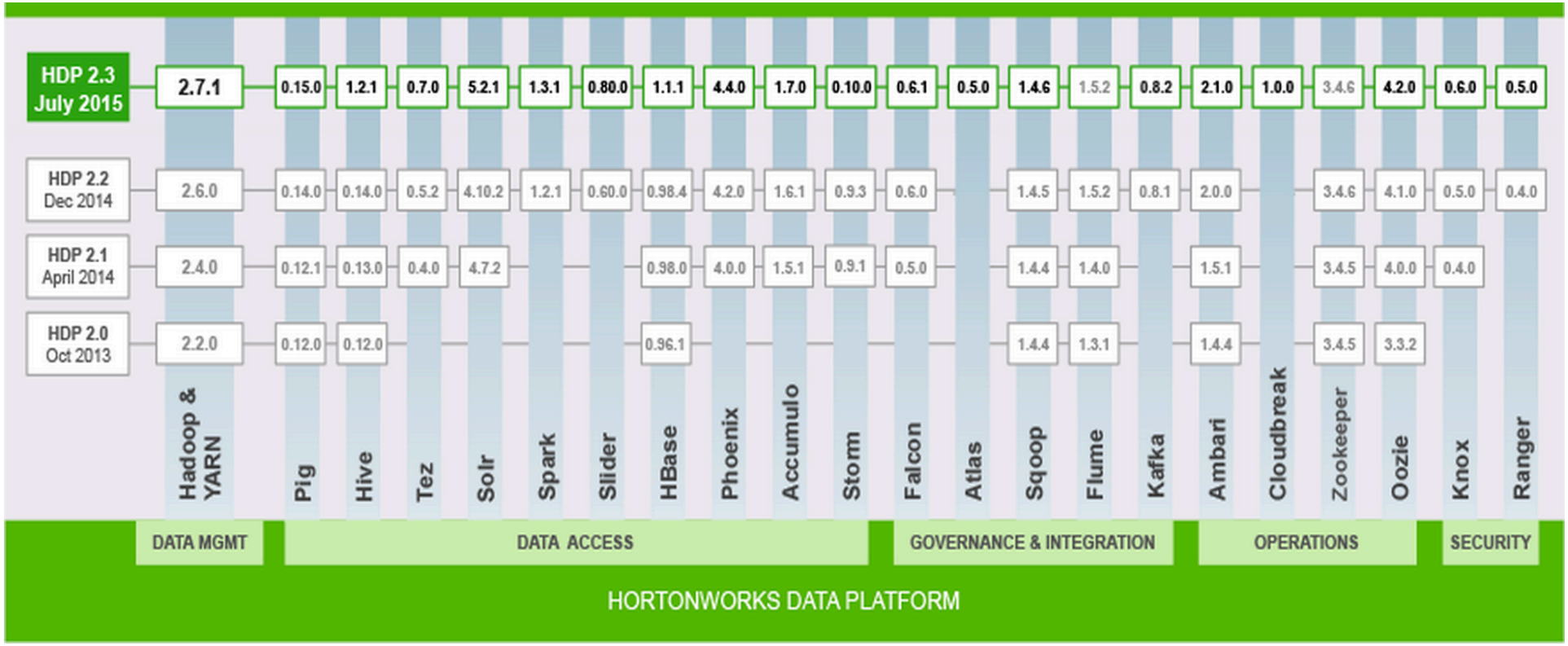
Reliable, scalable, real-time integration between Splunk and traditional relational databases

- Enrich search results with additional business context
- Easily import data into Splunk for deeper analysis
- Integrate multiple DBs concurrently
- Simple set-up, non-evasive and secure




.conf2015
Customer Open
Source Alternative
splunk>

Hadoop Ecosystem Options

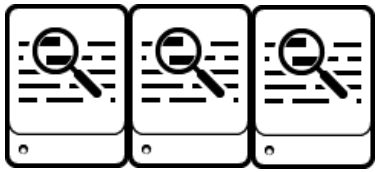


Hadoop Advantage / Disadvantage

Advantage	Disadvantage
Cheap Storage	Requires Coding for most Analytics
Batch Distributed Processing	No Visualization Tools
	No OOTB Apps / Solutions

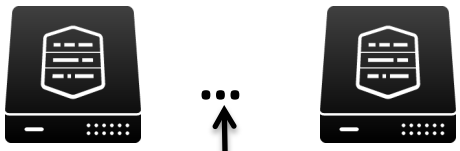

.conf2015
Real-World
Customer
Architecture
splunk>

Summary Architecture



3 instances
Splunk / Hunk / DB Connect
Search Heads

Real Time Data -
25 Indexers

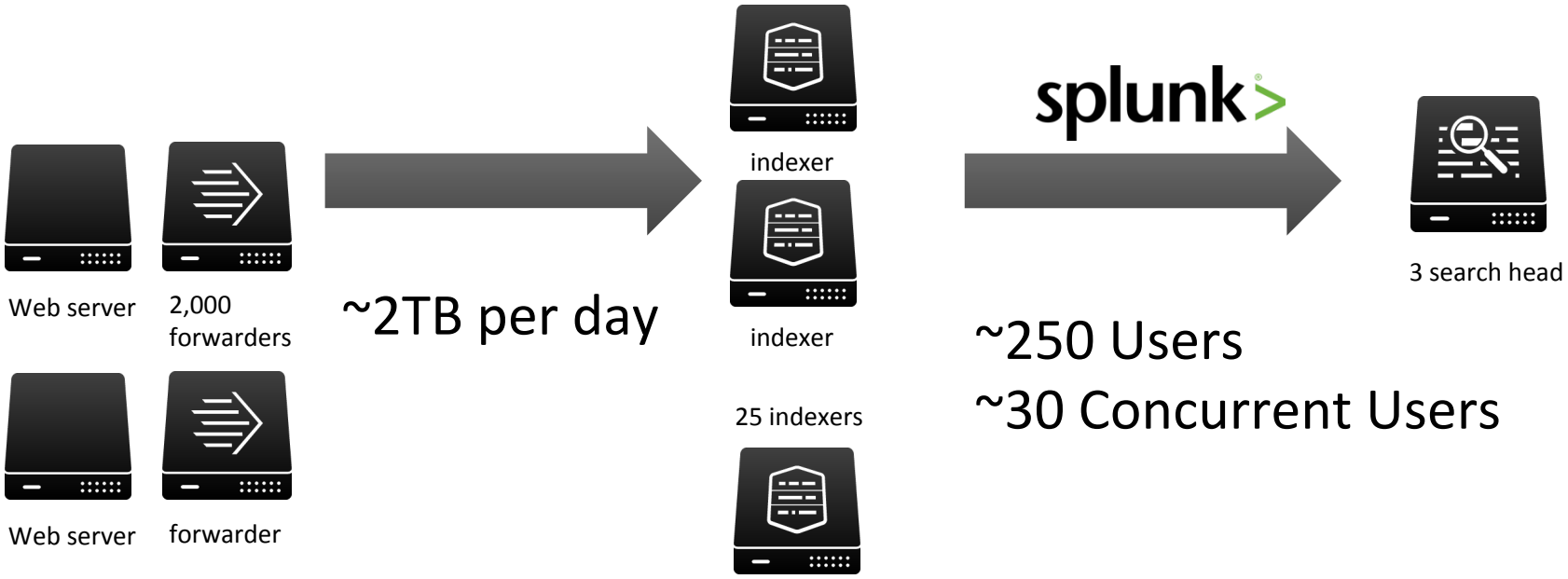


2000
Forwarders

Historical data (VIX)
60 Hortonworks nodes

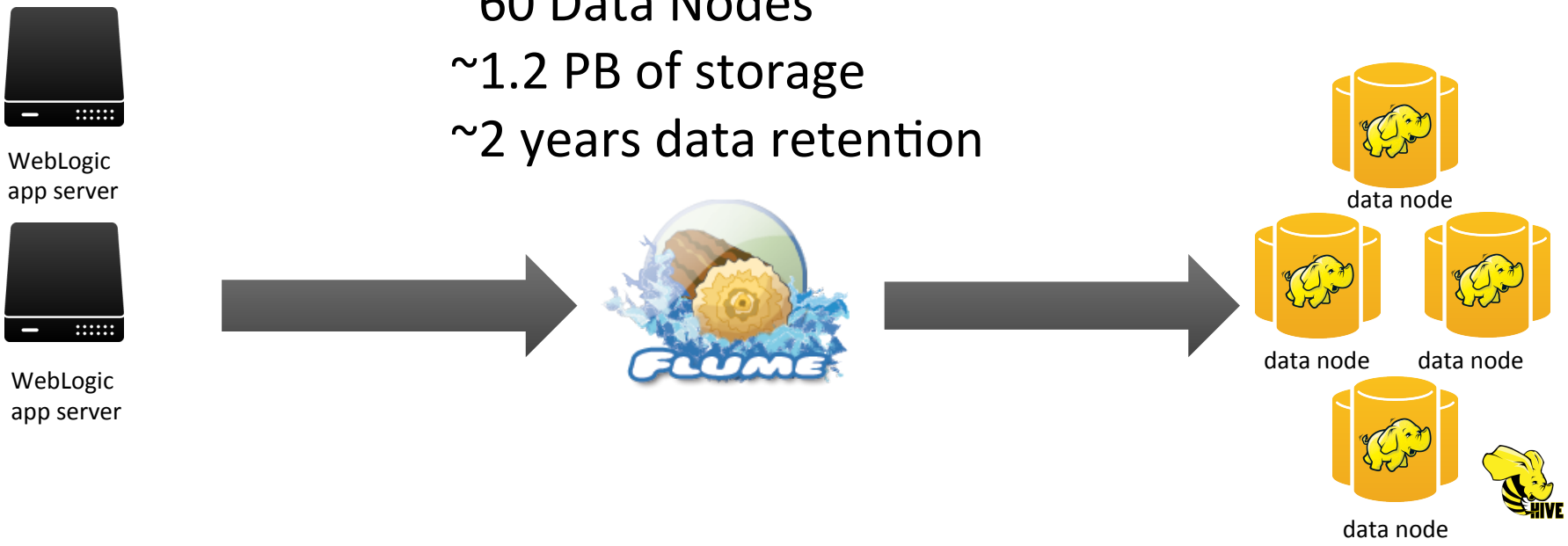
Enrichment data (lookup) -
MySQL DB

Splunk Deployment Architecture

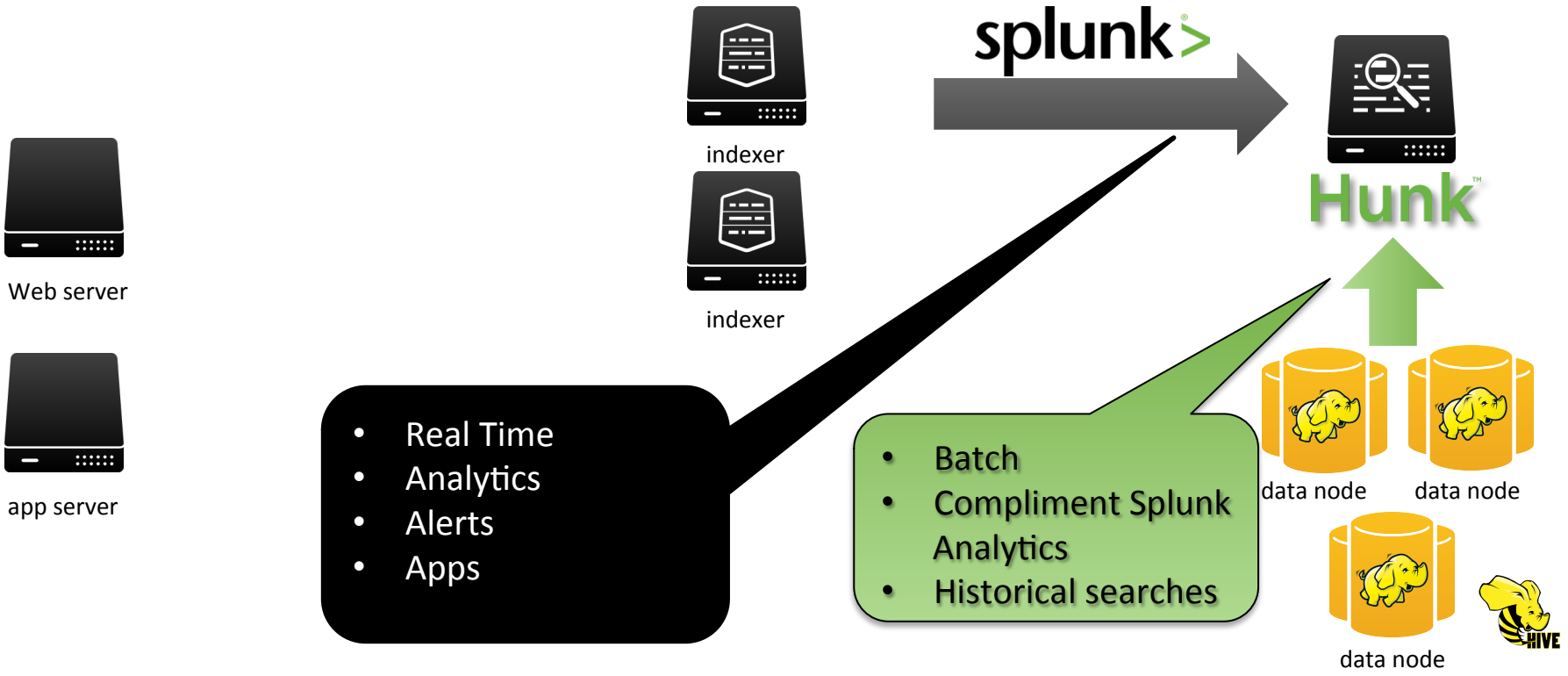


Hadoop Architecture

~30 Flume Agents
~60 Data Nodes
~1.2 PB of storage
~2 years data retention

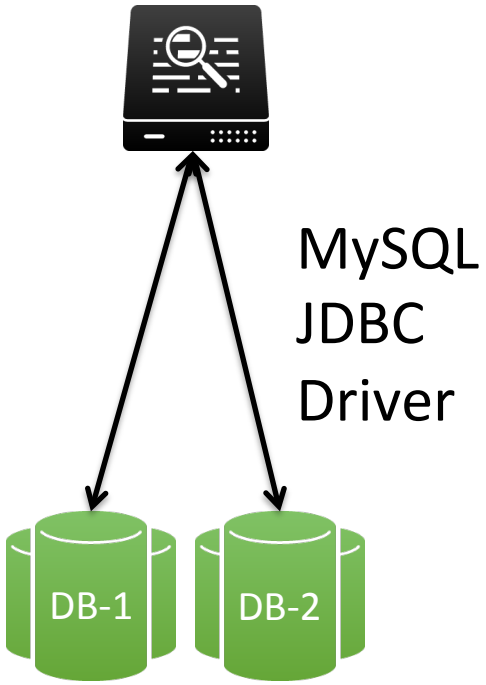


Splunk + Hunk = All the Data



DB Connect Architecture

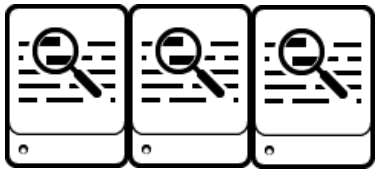
- Install DB Connect on a Search Head
- Use DB Connect for Lookup
- Several Lookups coming from two different MySQL Databases
- Lookup Enrich log data with business insight



DB - Architecture Performance Impact

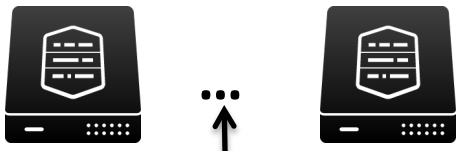
Command	Connection	Architecture
Indexing		
Inputs - dbmon-tail ** **Recommended**	Medium number of connections (Small amount of data - only delta)	DB to Index (connection pooling)
Inputs – dbmon-dump	Small amount of connections (Lots of data per connection)	DB to Index (connection pooling)
Outputs	Lots of DB Connections (Small amount of data)	Search Head to DB (connection pooling)
Not Indexing		
Search – DBXQuery	Lots of DB Connections	DB to Search Head
Lookups ** **Selected this option**	Lots of DB Connections	DB to Search Head

Summary Architecture



3 instances
Splunk / Hunk / DB Connect
Search Heads

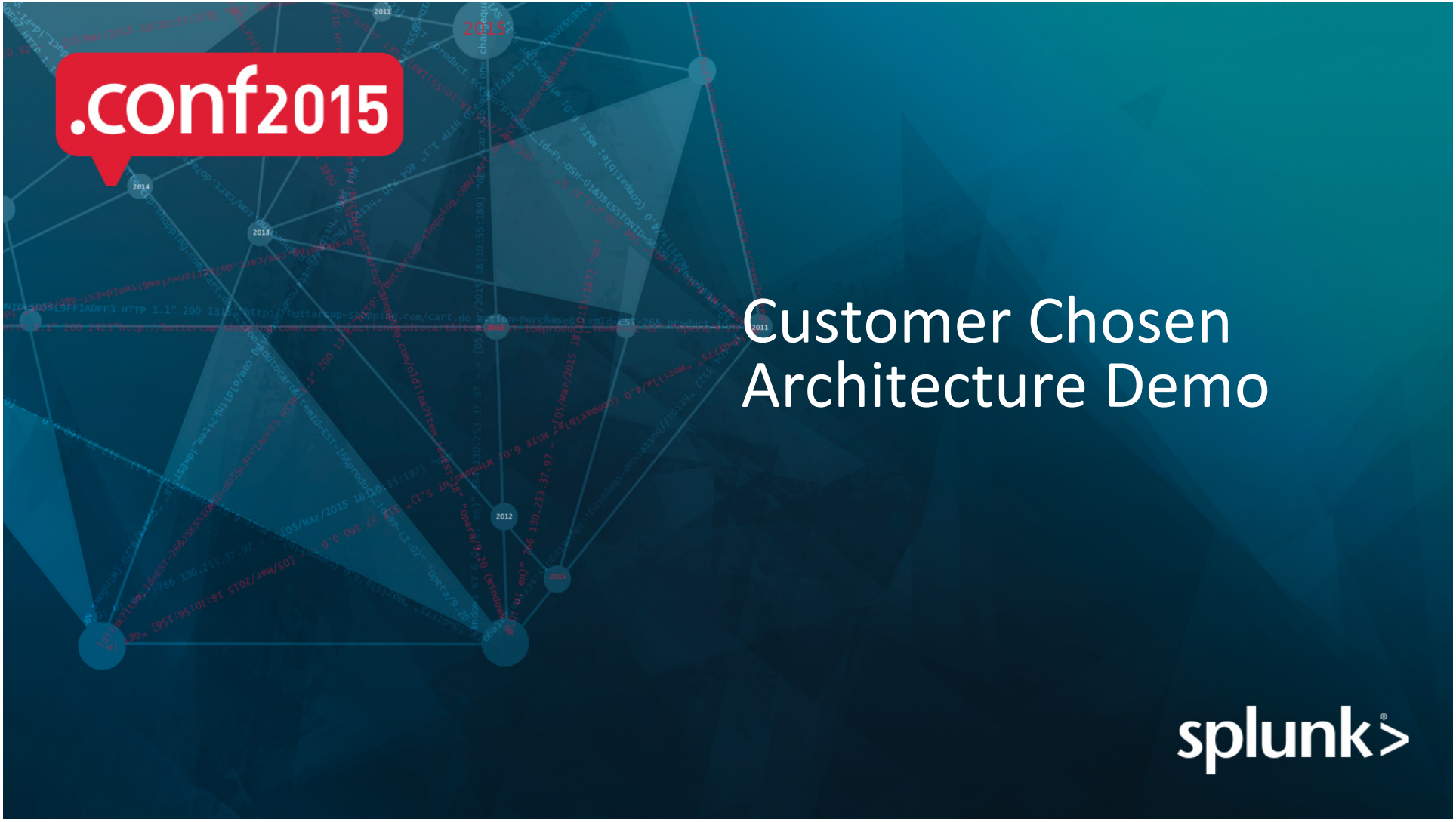
Real Time Data -
25 Indexers



2000
Forwarders

Historical data (VIX)
60 Hortonworks nodes

Enrichment data (lookup) -
MySQL DB


.conf2015
Customer Chosen
Architecture Demo
splunk>

THANK YOU

splunk>