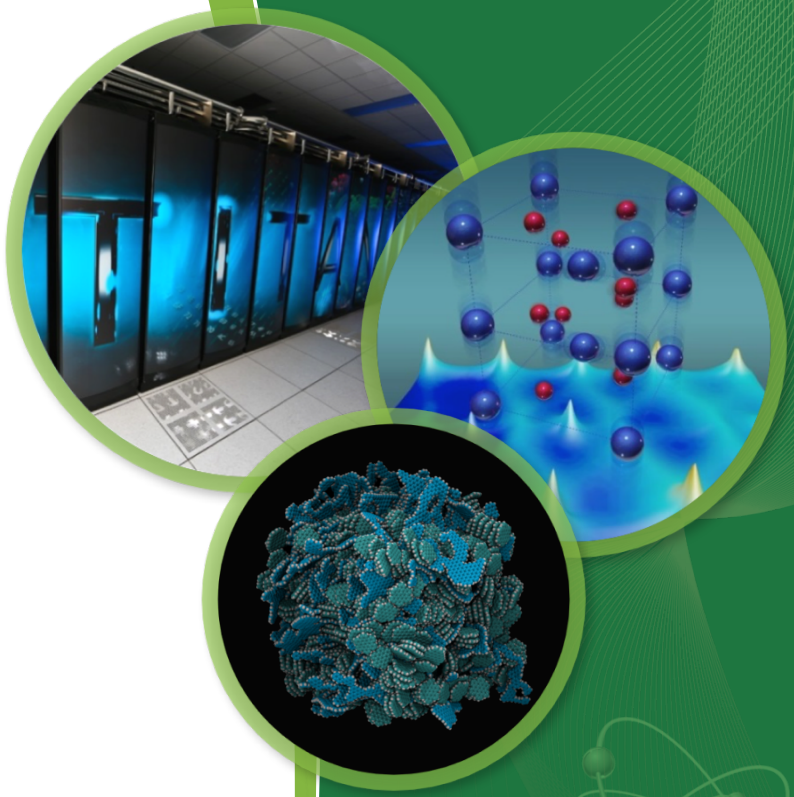
Big Data Meets HPC Log Analytics: Scalable Approach to Understanding Systems at Extreme Scale



Byung Hoon Park
Saurabh Hukerikar
Ryan Adamson, and
Christian Engelmann

Oak Ridge National Laboratory

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

Background

- HPC Reliability, Availability, and Serviceability (RAS) Logs are valuable resources to obtain clues for system status, failures (or system anomalies), performance degradation, etc.
- Many researchers want access to logs, However …
  - Data are often heterogeneous, unstructured, huge in volume, ant not stored in database for retrieval.
- Capability of storing, analyzing, and visualizing RAS logs is greatly desired.
- Our efforts to create ***Big Data*** based HPC log analytic framework for a large pool of users is presented.

OAK RIDGE National Laboratory

HPC Logs and Big Data

"***Big data*** *is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include* ***capturing data***, ***data storage, data analysis***, ***search***, ***sharing***, ***transfer***, ***visualization***, ***querying***, *updating and information privacy.*"

- Wikipedia

OAK RIDGE National Laboratory

Overall Roadmap to HPC Log Analytics

Database Selection and Design of Data Model

- Flexible and Time series friendly data model for HPC event data

Data Transformation and Ingest into Database

- Event Definition and ETL Process

Selection of Computational Model

- Identification of the best computational model for the data model

Offline Analytics Framework

- Serves analytics on historical data

Online Analytics Framework

- Real-time monitoring and analytics

K RIDGE
nal Laboratory

Database Selection for HPC Log Events

- Scalability
  - Database should be capable of accommodate ever-increasing volume of data
- Low latency and High availability
  - Should serve interactive analytics and near real-time query requests.
  - Should survive failure.
- Flexibility
  - Should represent various types of events
- Time series friendly
  - Time series analytics over an arbitrary time interval

OAK RIDGE National Laboratory

Database Selection

Relational DB

Document

Key Value Store

Column Store

Graph DB

…



OAK RIDGE National Laboratory

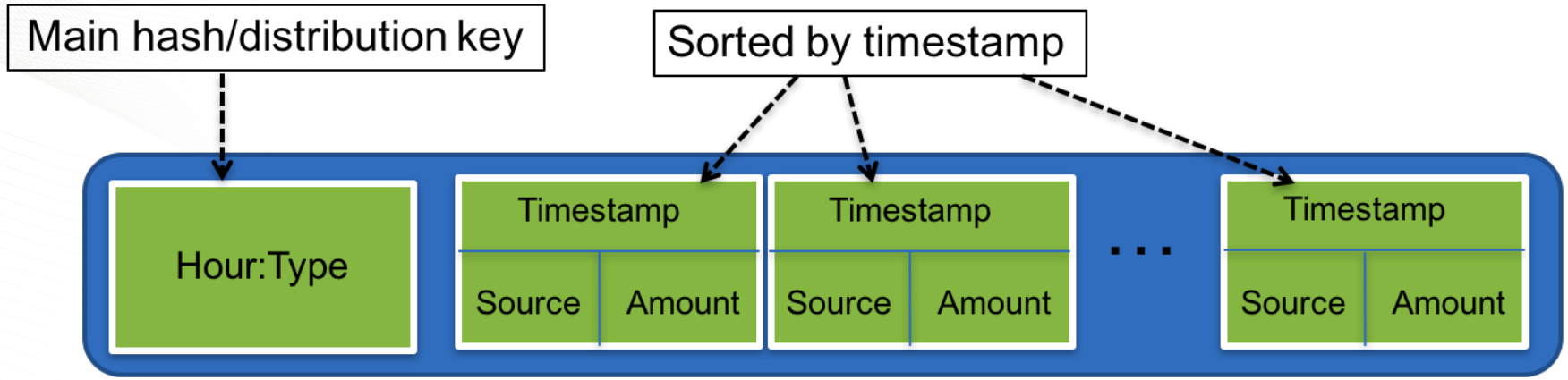
NoSQL Database

TYPES	PERFORMANCE	SCALABILITY	FLEXIBILITY
KEY-VALUE STORE	high	high	high
COLUMN STORE	high	high	moderate
DOCUMENT	high	moderate	high
GRAPH DATABASE	moderate	moderate	high

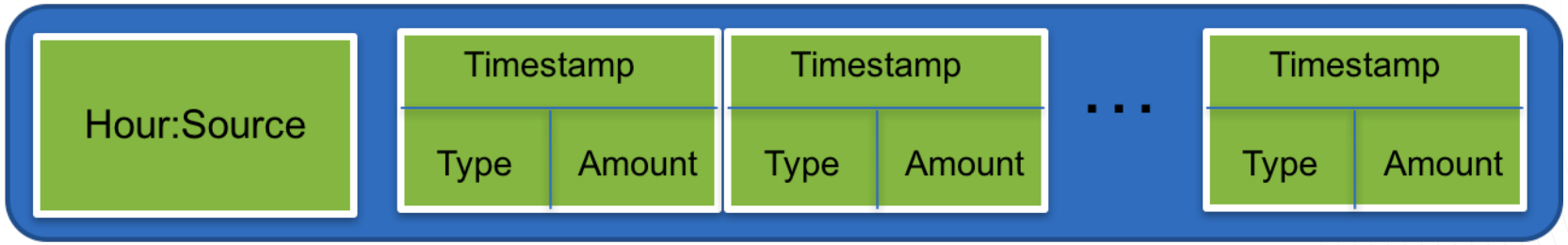
Cassandra
- Column family based key value store (flexible)
- Hash-based distributed data model (scalable)
- Convenient to store data in a time series
- Shows superior performance over other NoSQL DBs

OAK RIDGE National Laboratory

Event Data Model with Redundancy for Efficient Retrieval



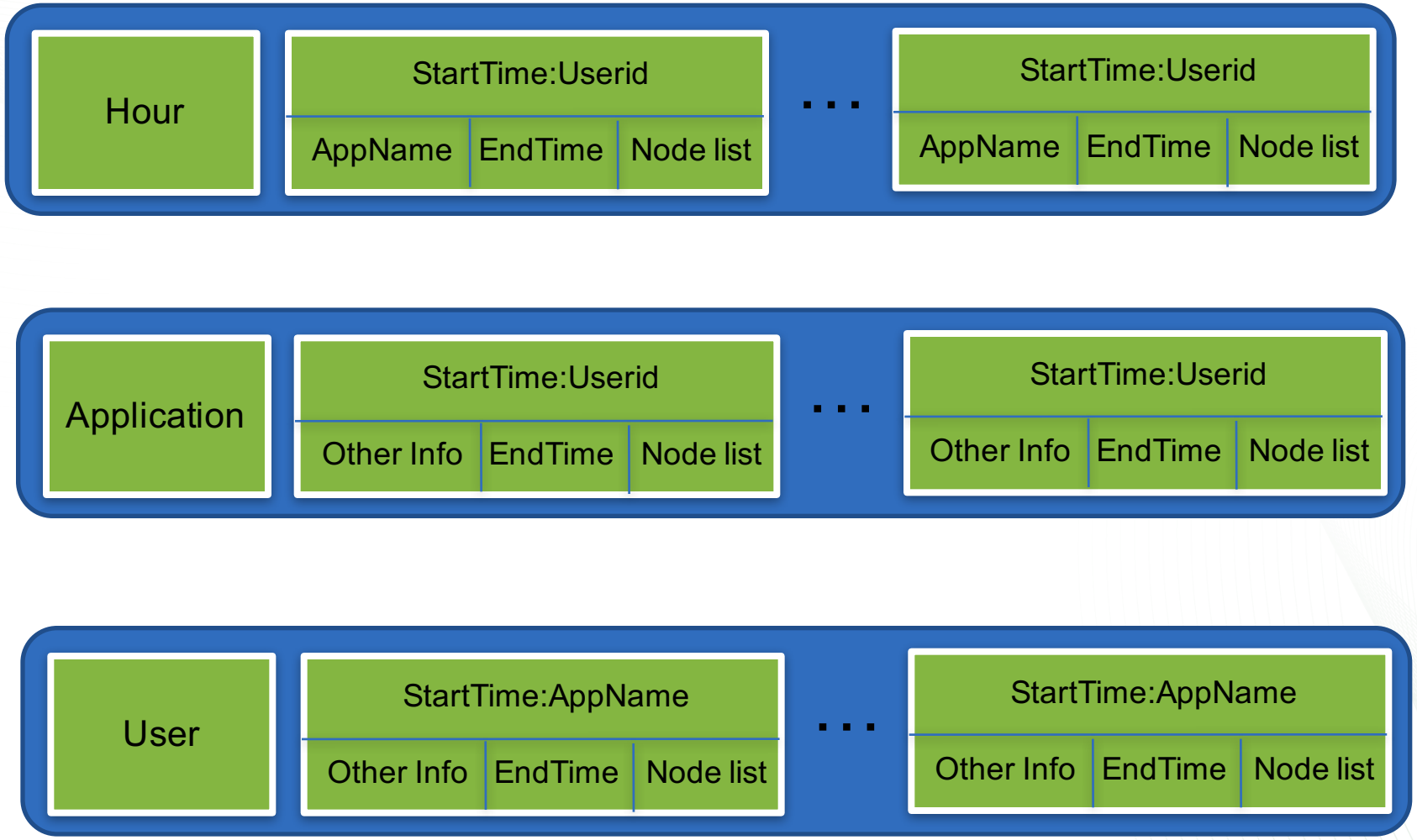
A data row for event occurrences arranged by event type



A data row for event occurrences arranged by location of occurrence

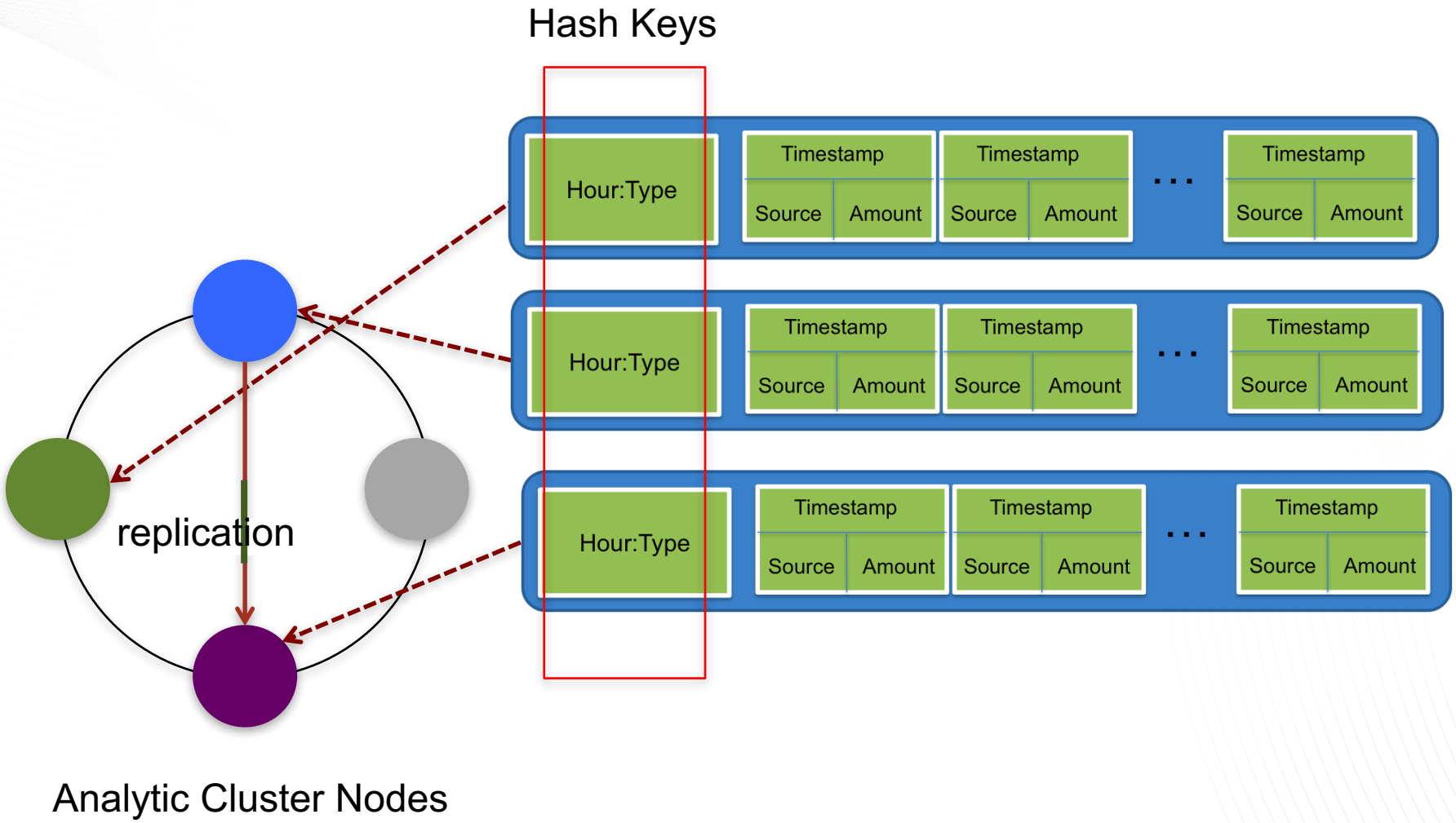
OAK RIDGE National Laboratory

Workload Data Model with Redundancy



OAK RIDGE National Laboratory

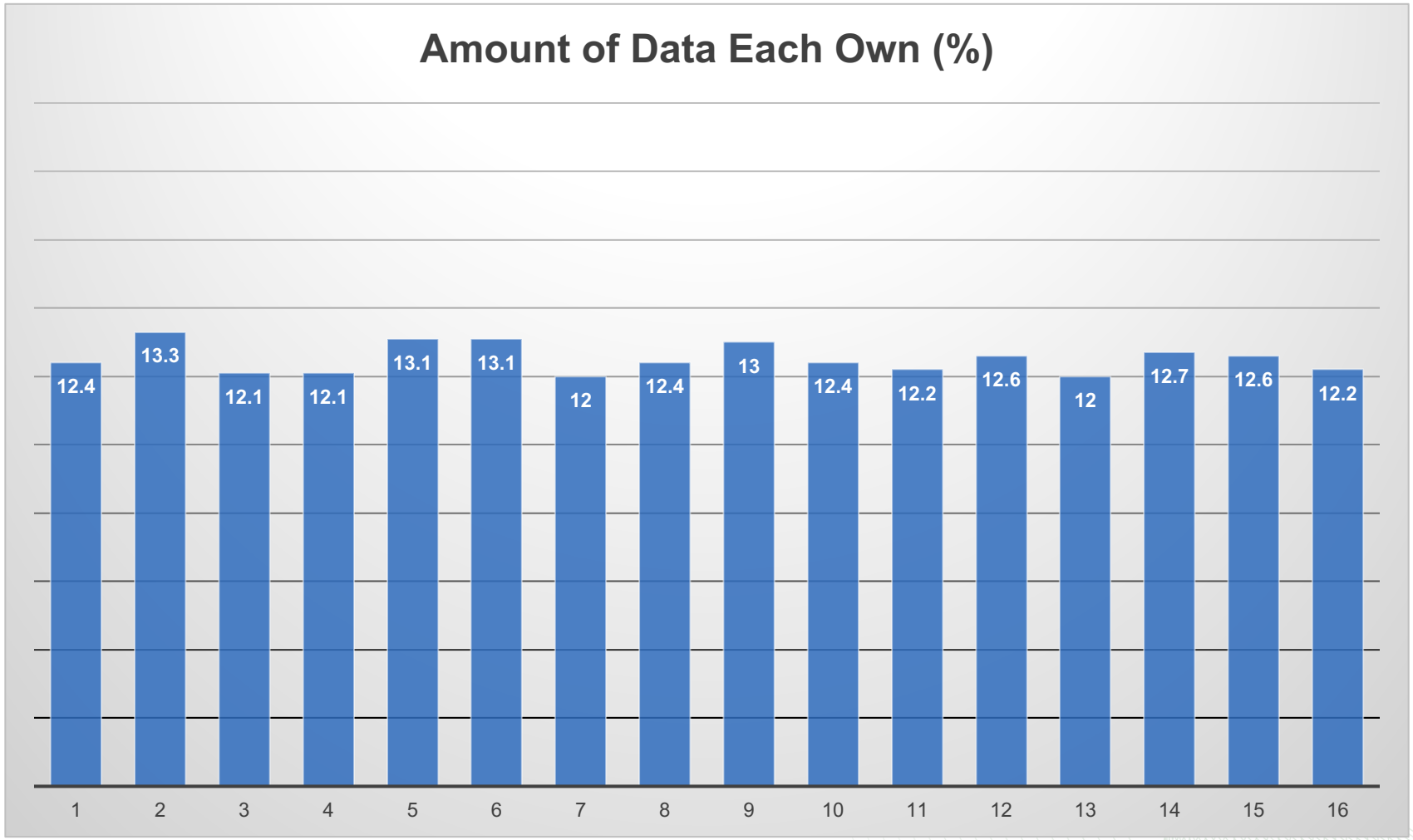
Data Tables are Distributed by Hash Keys



OAK RIDGE National Laboratory

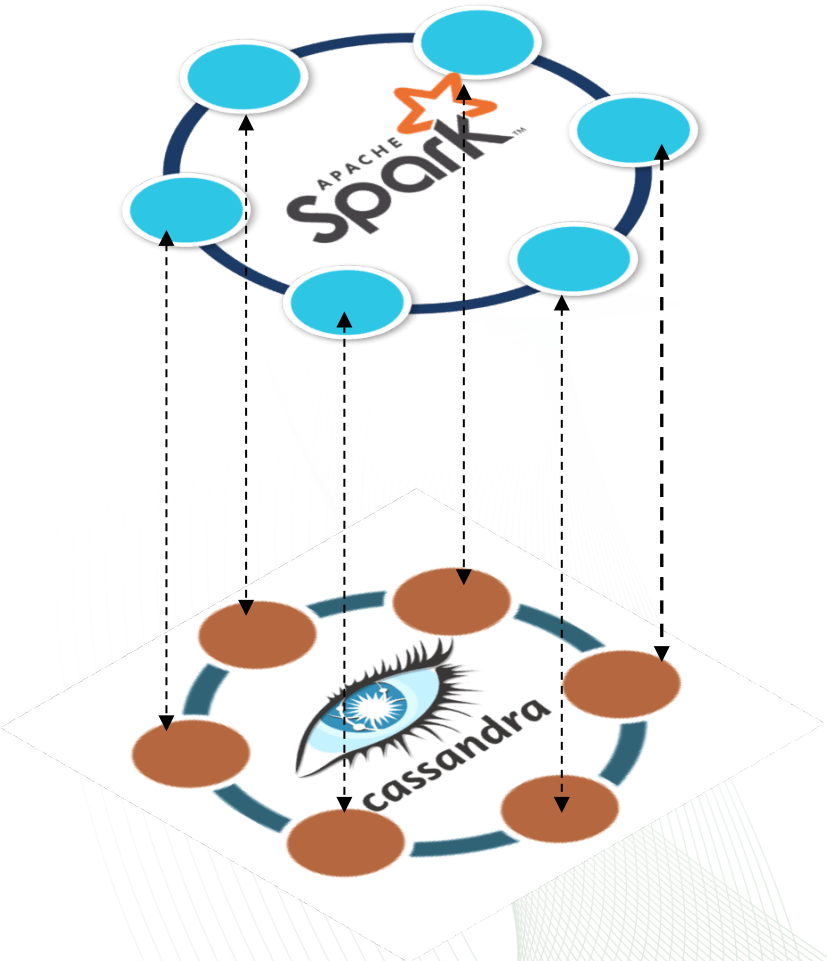
Data Distribution over Cassandra Nodes

- A month of data are populated over 16 node cluster.
- Percentages of data owned by each Cassandra node are measured.
- The result shows our schema design results in a good distribution.



Spark: Computation Model for Cassandra

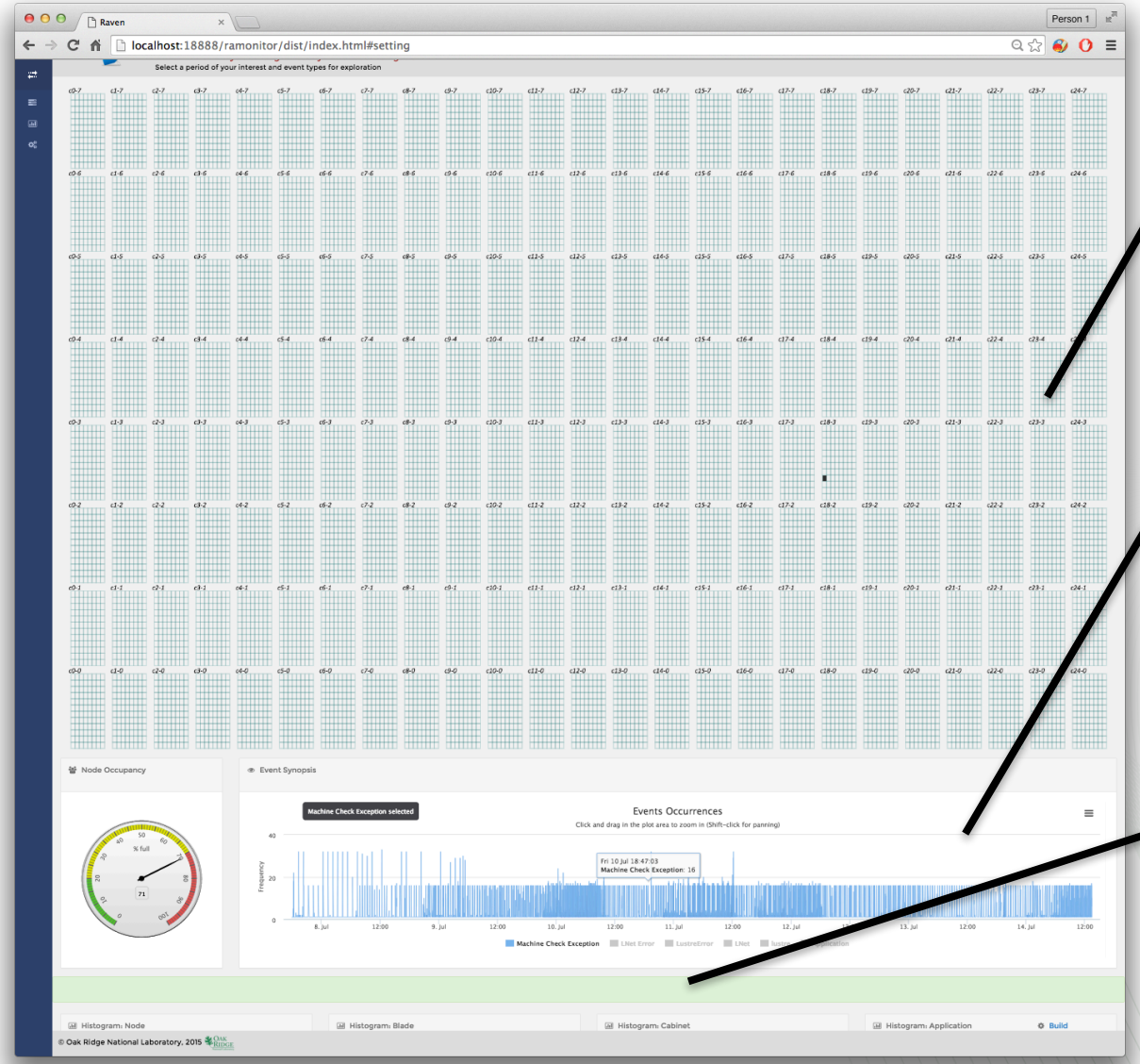
- Spark is a distributed in-memory map/reduce computation framework.
- Spark includes a rich suite of analytic modules
  - Statistics, machine learning, stream mining, text mining, etc.
- For log analytics, data in the Cassandra cluster need to be moved into the distributed memory.
- To maximize locality, a spark node is installed in each Cassandra node.



OAK RIDGE National Laboratory

The Frontend User Interface

- Users interacts with the framework by creating a query of *task* and *context*.
- Context is:
  - Event type
  - Application
  - Location
  - User
  - Time period
  - Or combination of all above
- Frontend provides interfaces to create a context.



Physical Map

Time Interval

Event/Application Selection

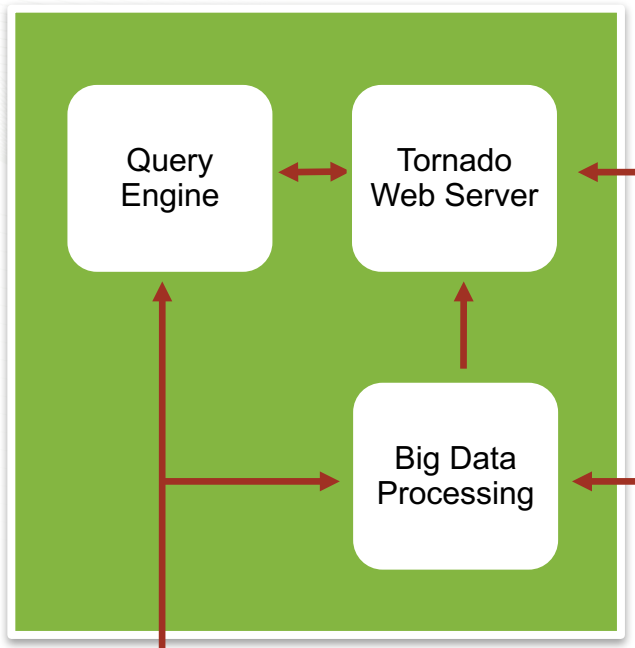
OAK RIDGE National Laboratory

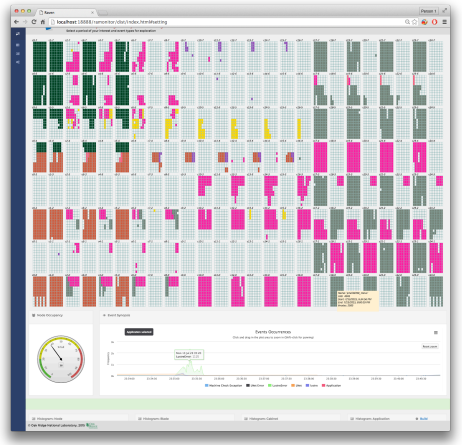
Offline Analytics Framework

- Cassandra + Spark Cluster as data and computation component
  - Increased performance thanks to data locality
- Analytic server for multitudes of users
  - Query processing engine
  - Web-based frontend
- Frontend as an event browser offering basic analytic capabilities
- Enabling accessibility by providing REST server
  - Users can write/execute their own codes for analytics

OAK RIDGE National Laboratory

Offline Analytics Framework Architecture





Frontend





Spark Job Server

REST



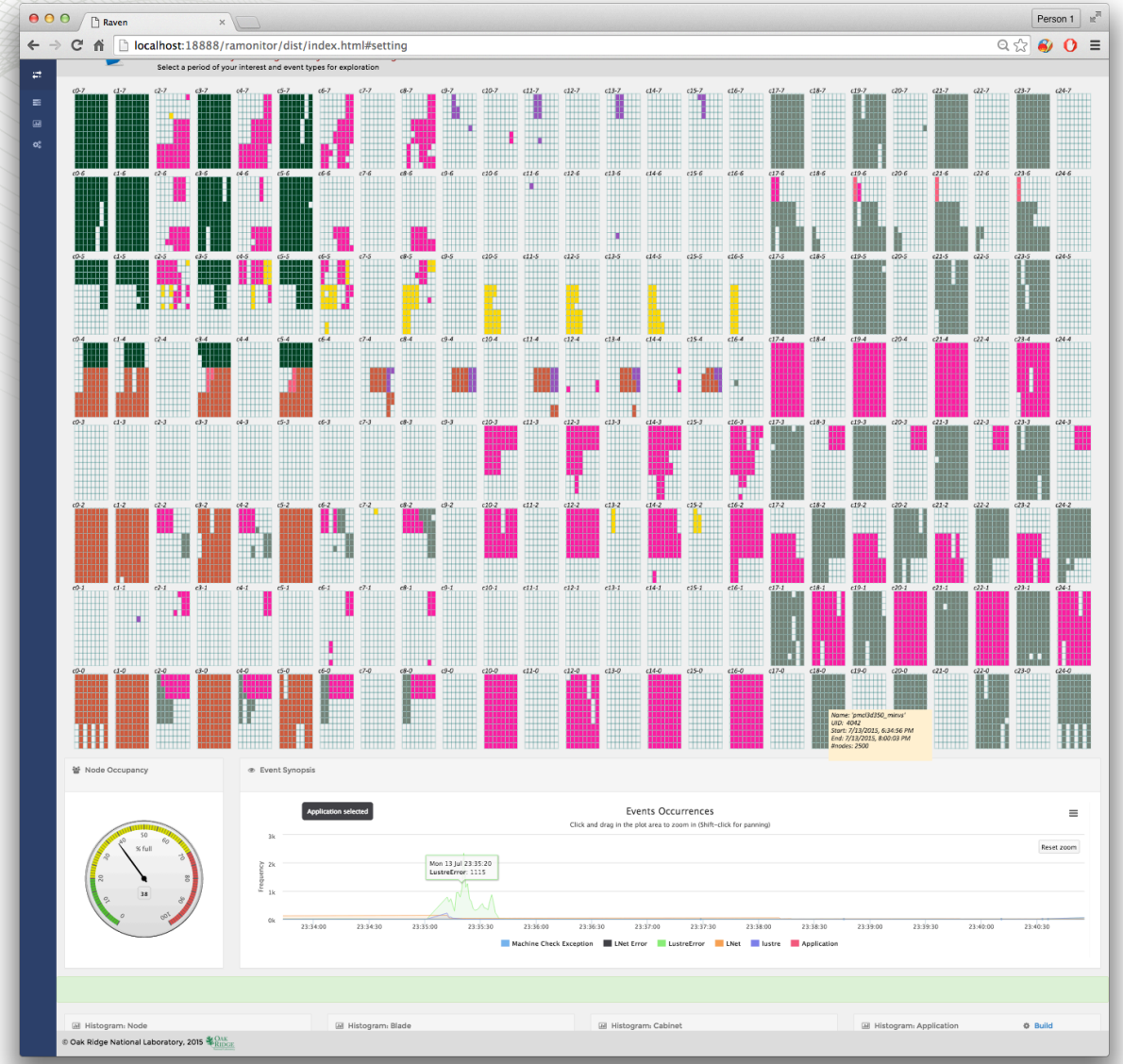
Java

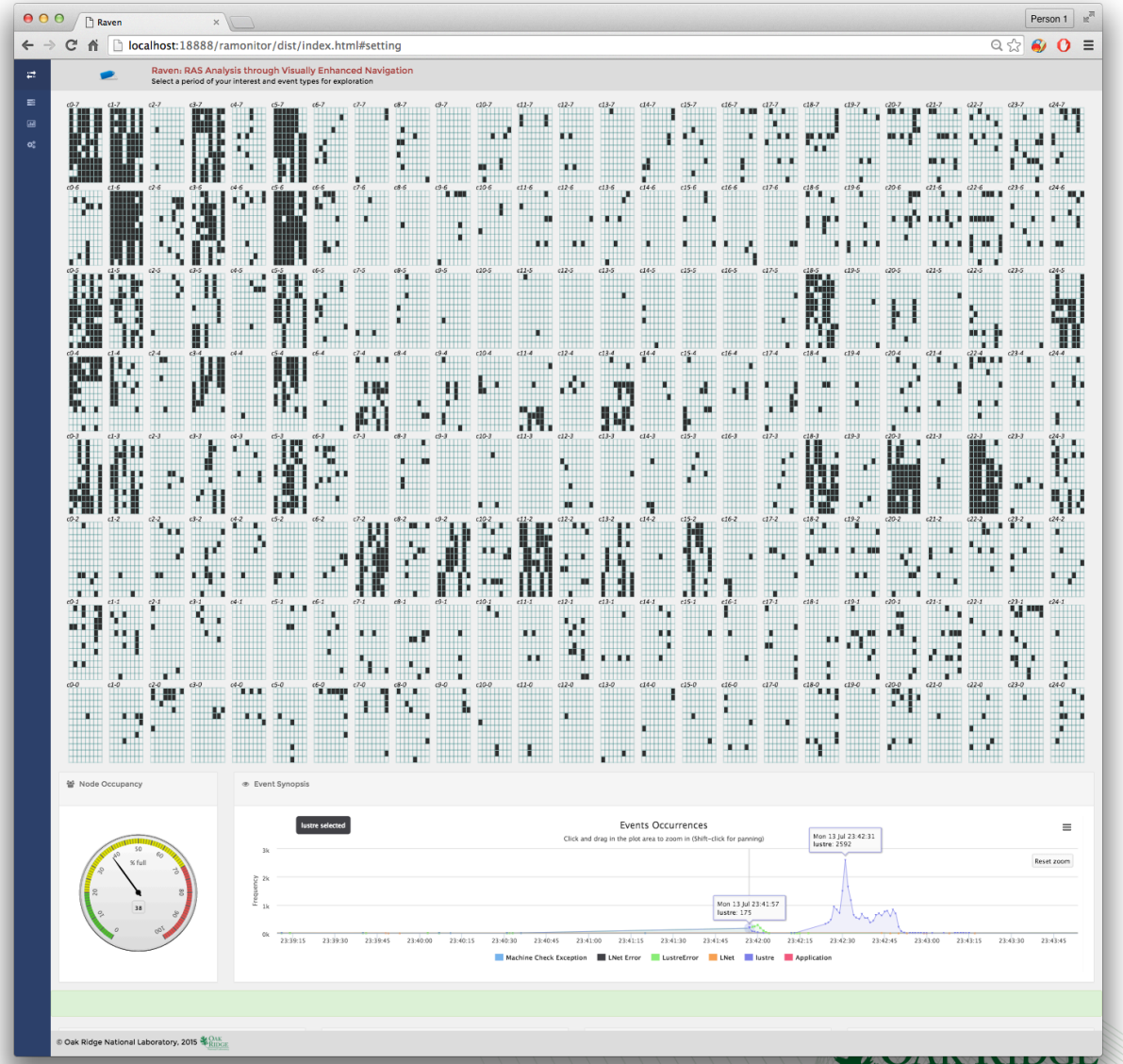
Scala

Python

OAK RIDGE National Laboratory

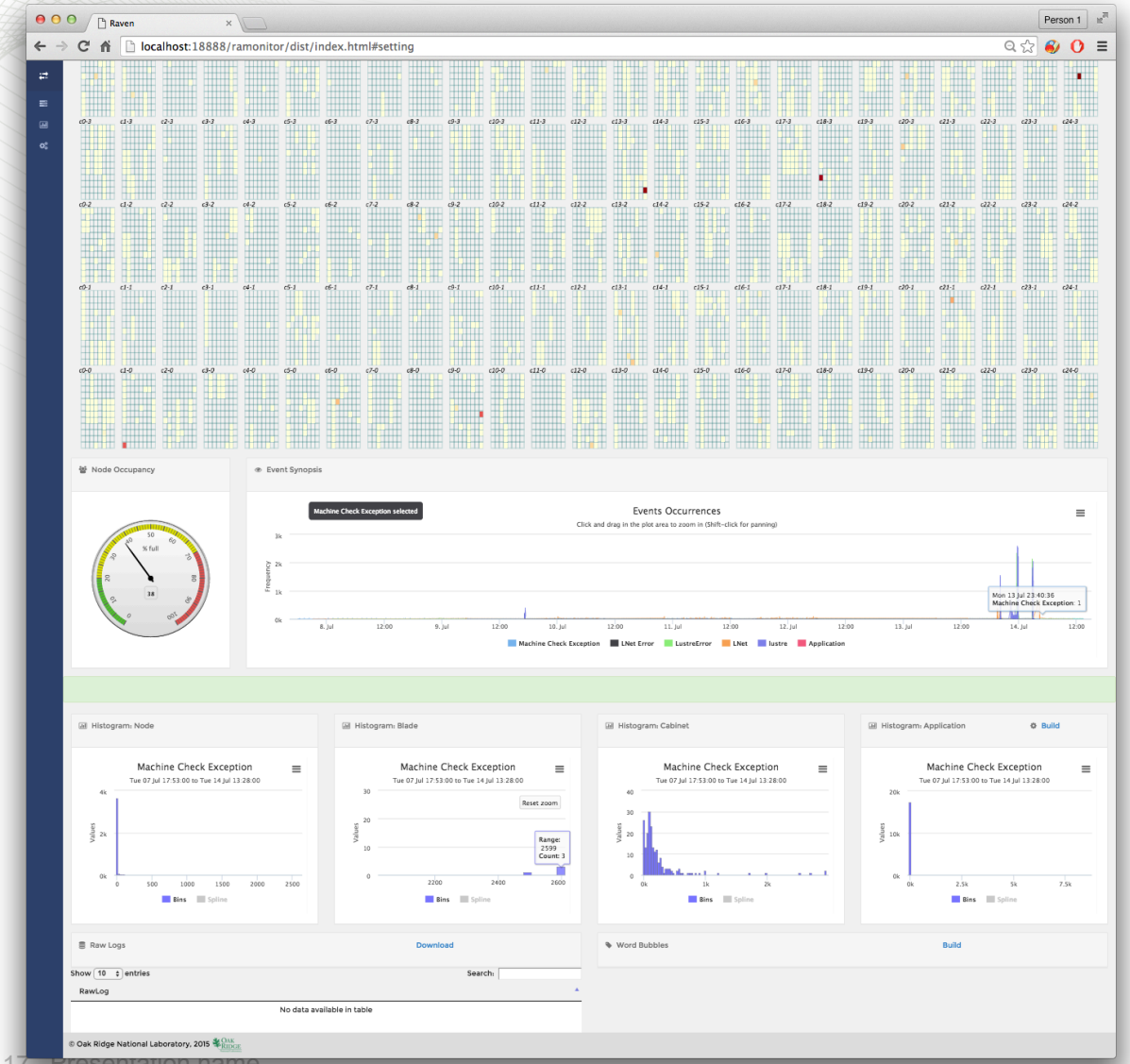
Spatiotemporal Browsing of Events and Applications

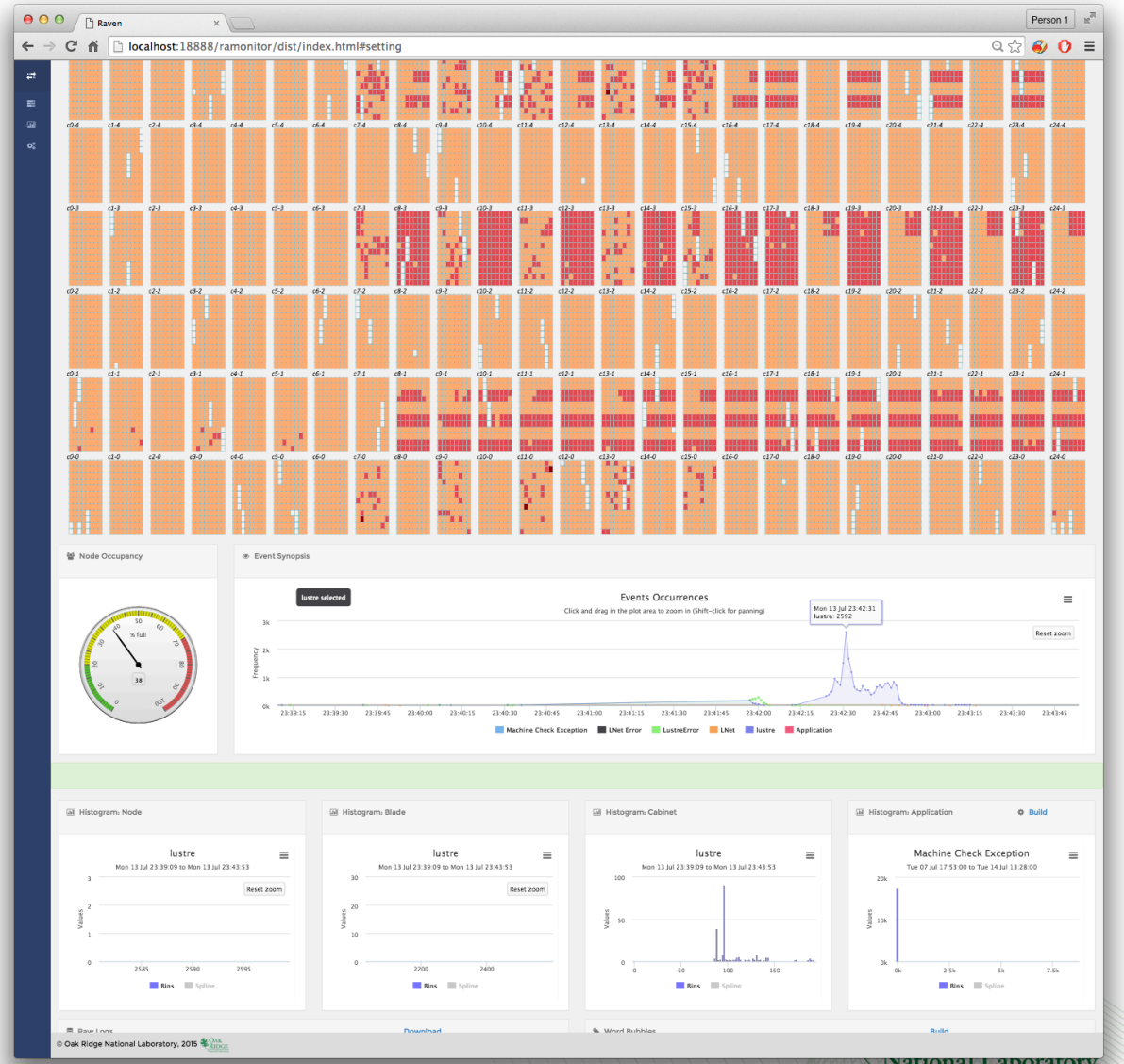




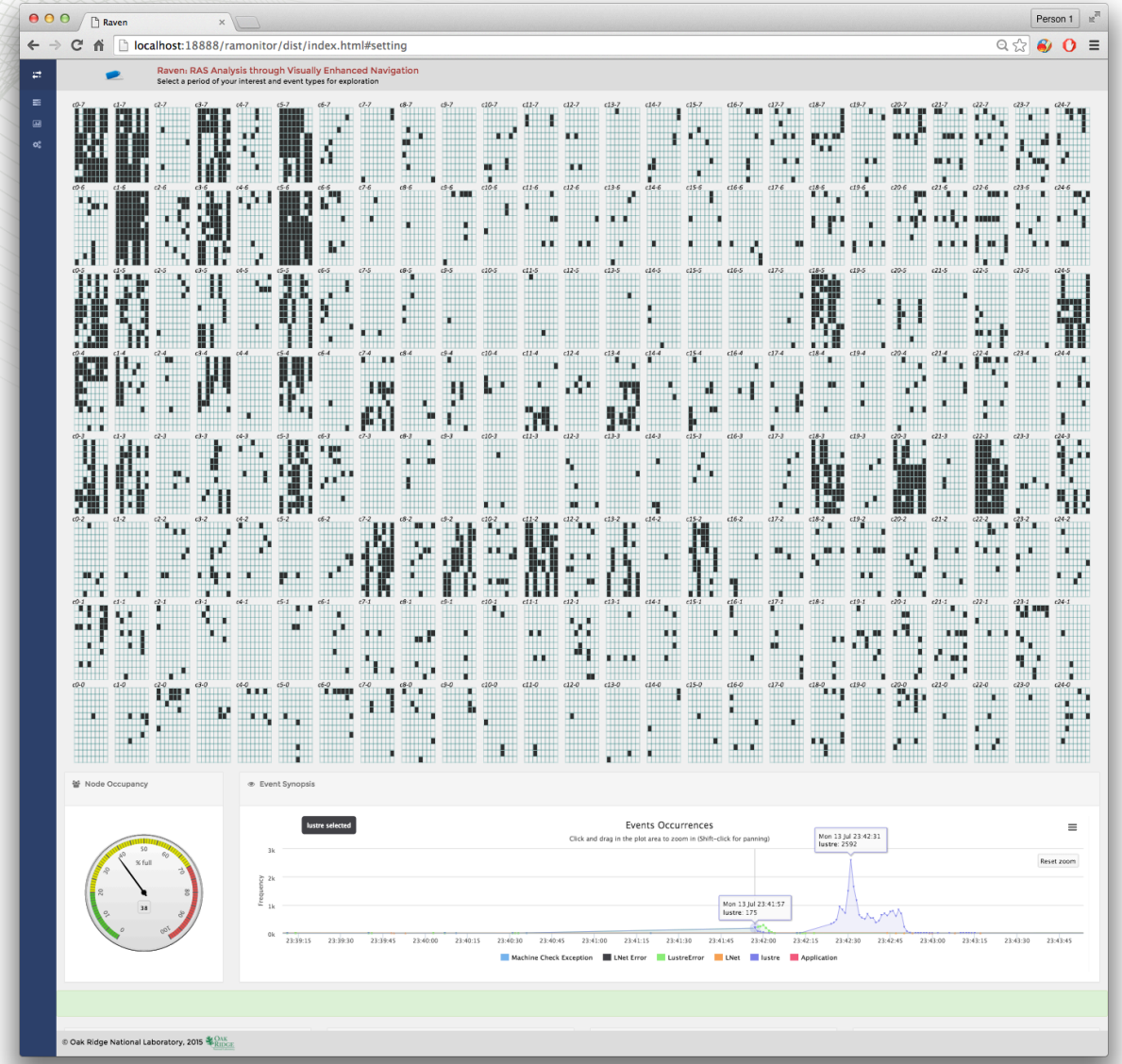
National Laboratory

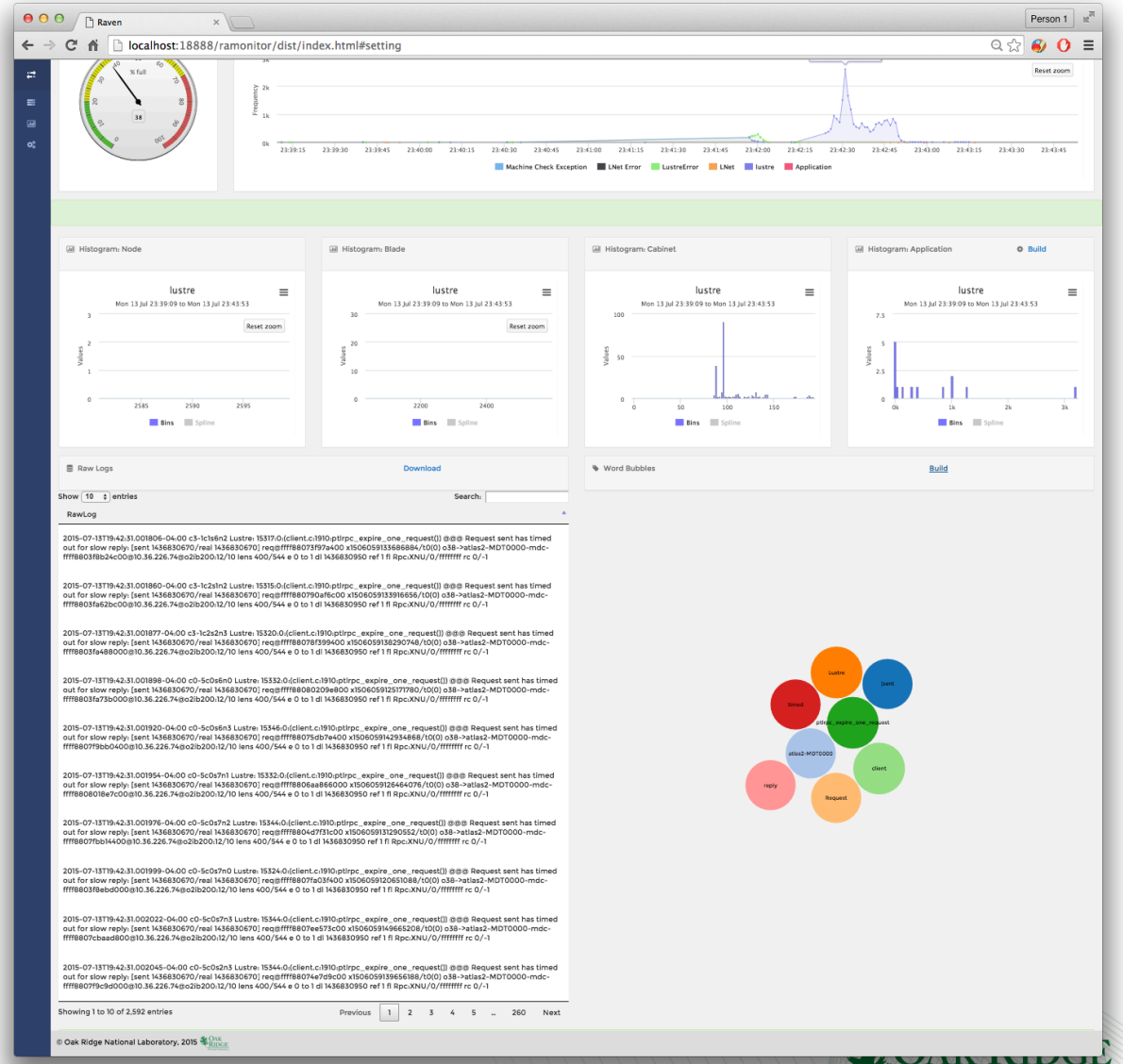
Statistics Computation using Spark





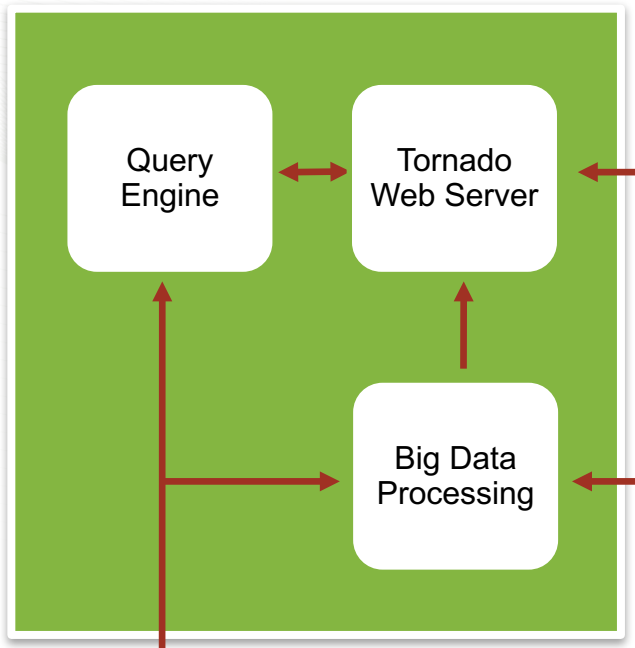
Simple Text Mining using Spark

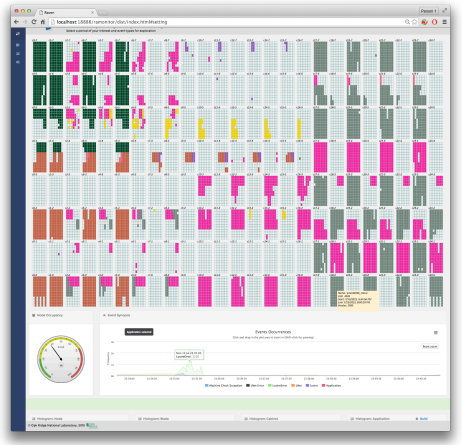




OAK RIDGE National Laboratory

Offline Analytics Framework Architecture





Frontend





Spark Job Server

REST



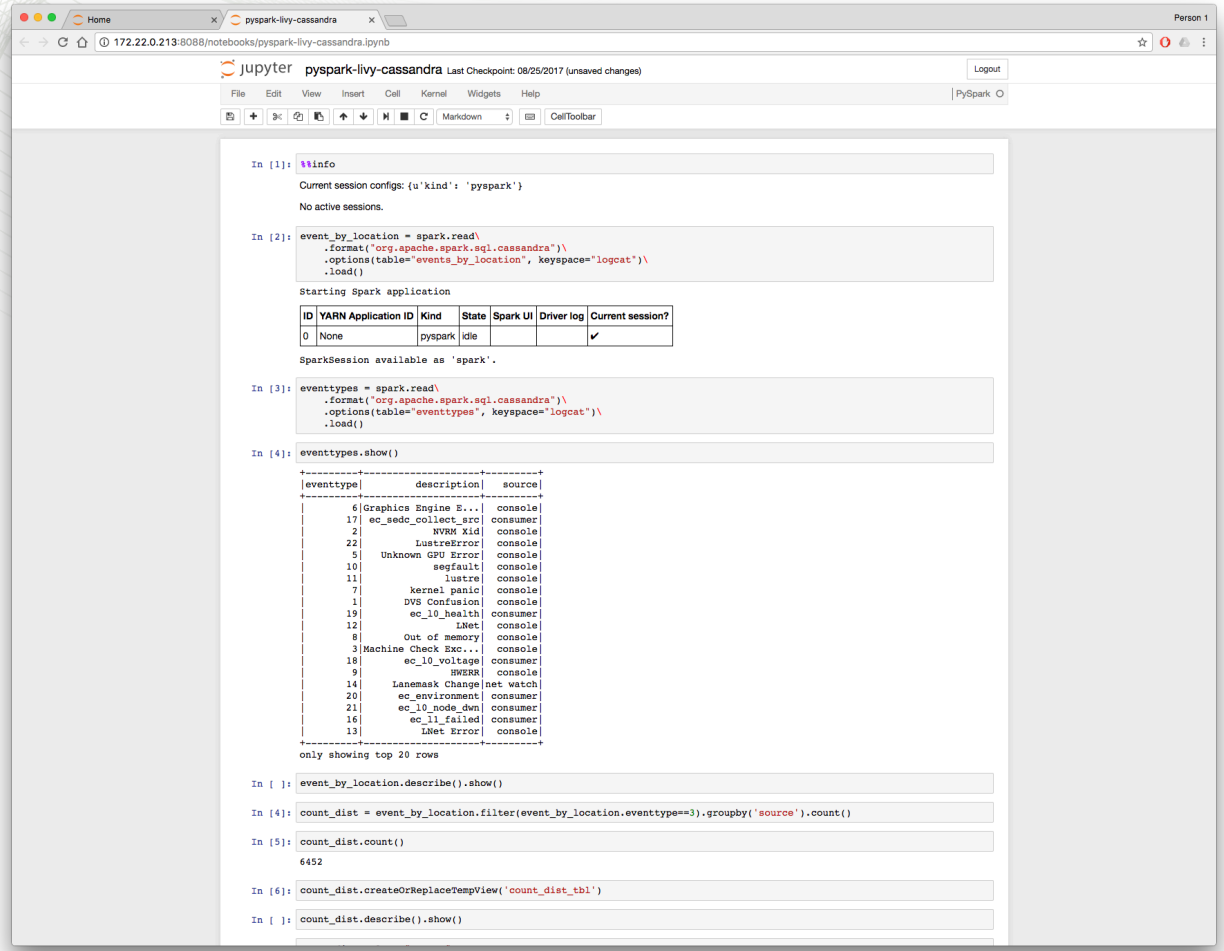
Java

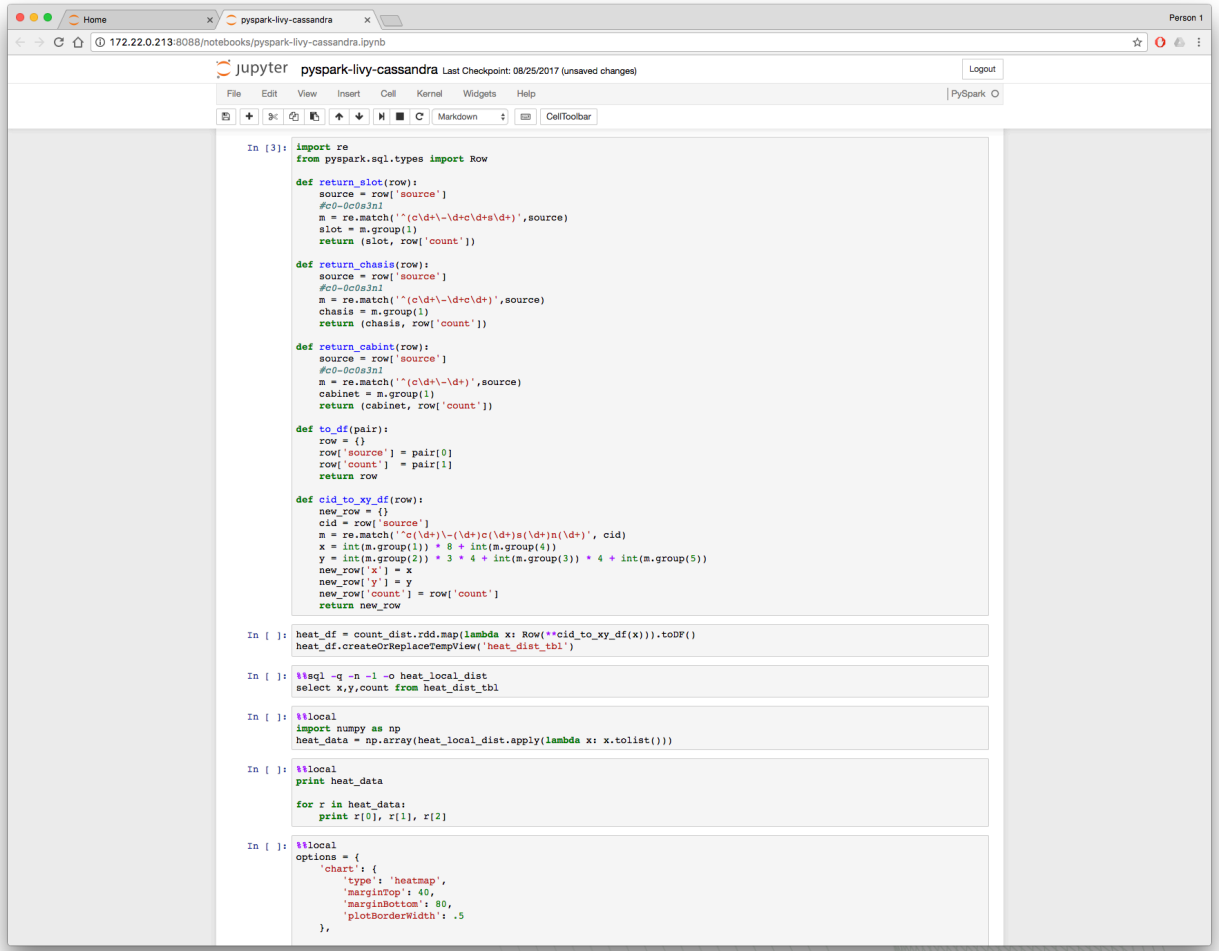
Scala

Python

OAK RIDGE National Laboratory

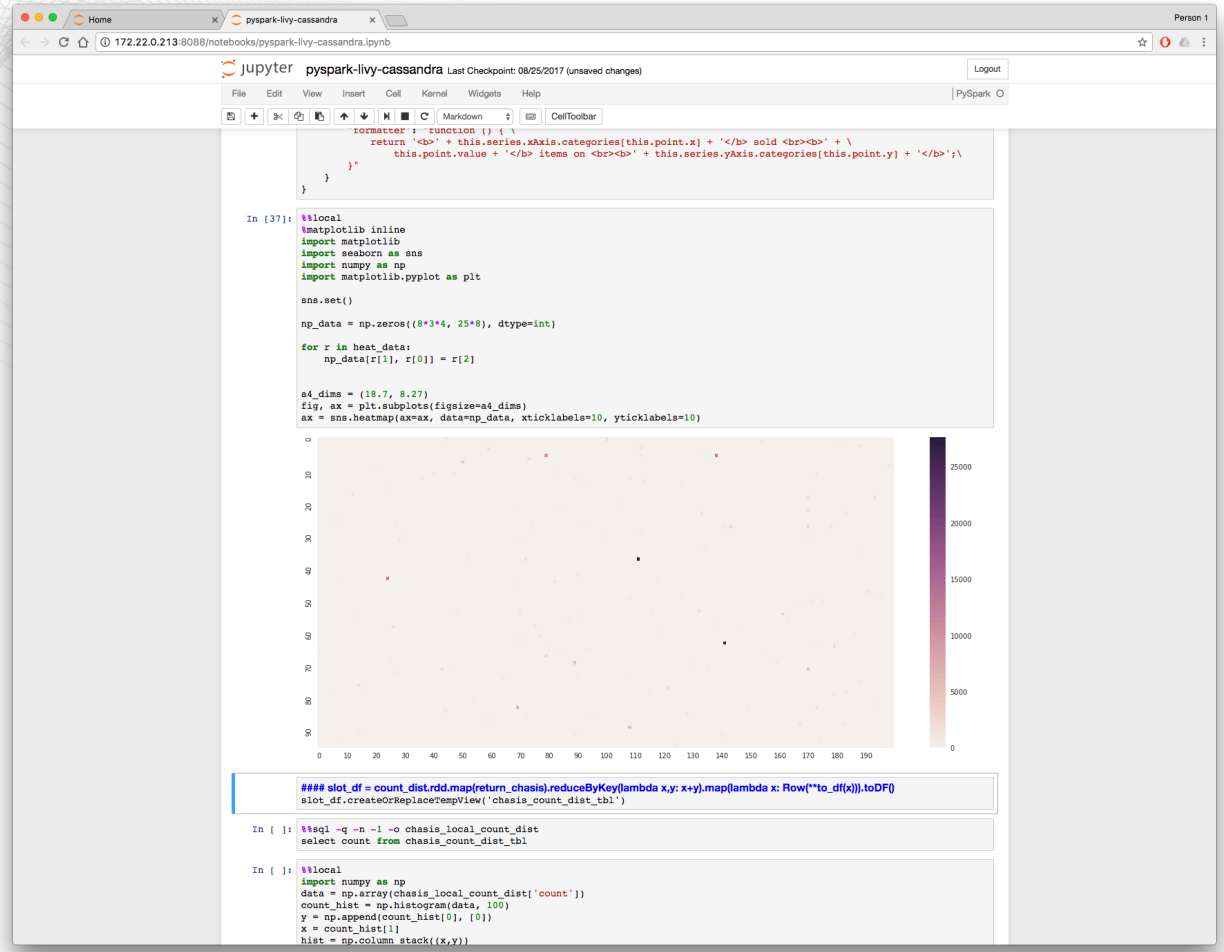
Jupyter Client for Ad Hoc Analysis

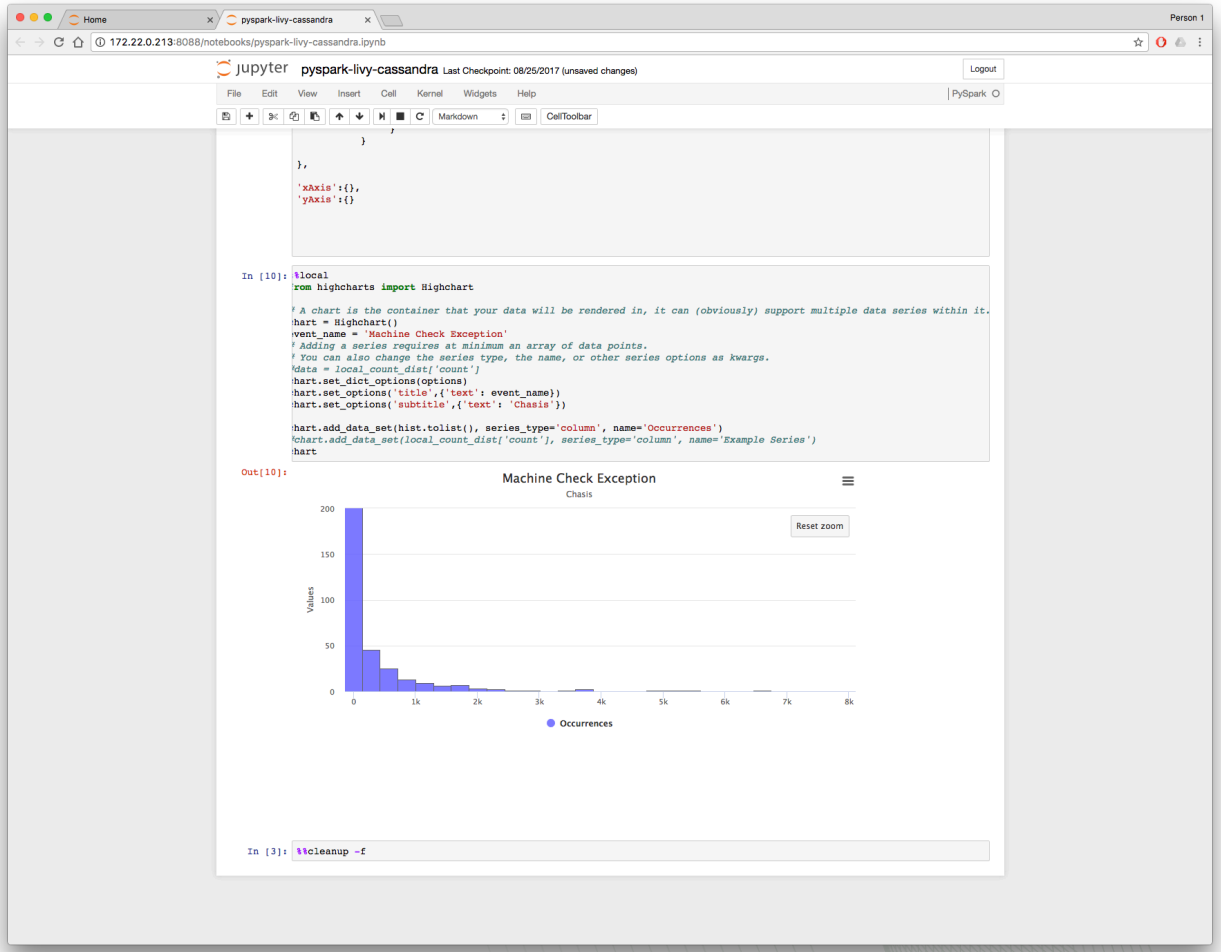




OAK RIDGE National Laboratory

Jupyter Client for Ad Hoc Analysis Cont.





OAK RIDGE National Laboratory

Online Monitoring & Analytics

- Facilitates online analytics
  - Real time failure detection from event streams
  - Maintaining up-to-date statistics regarding system usages, event occurrences, application runs, etc.
  - Providing logs as multiple time series streams of events
- Challenge is to bring processed data to end user with *negligible* lag
  - Data co-location
  - Processing of fast stream of large volume
- Our approach
  - Publish events from where they are generated
  - Consume events in parallel using Spark

OAK RIDGE National Laboratory

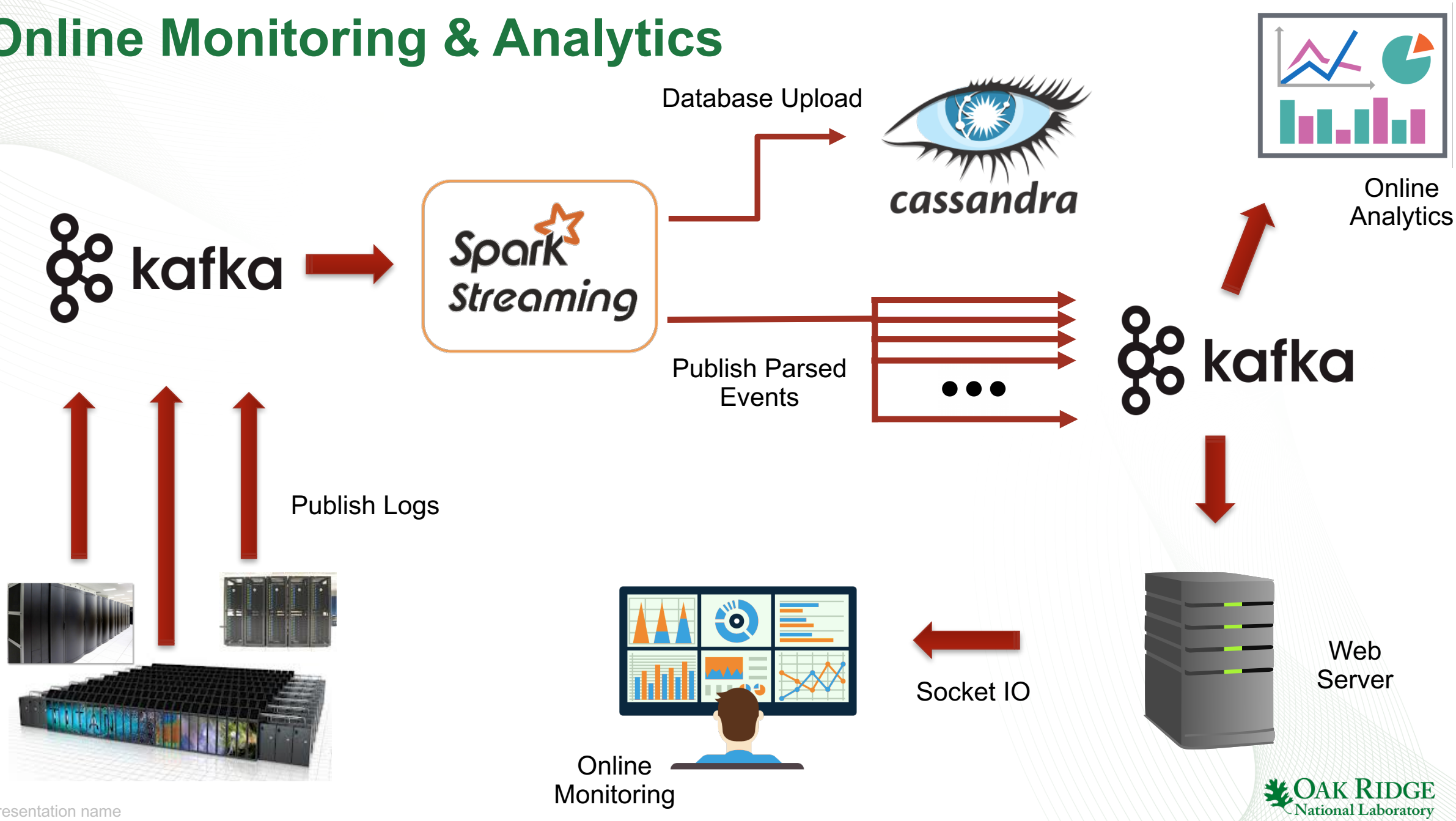
Online Monitoring & Analytics Cont.

- Oak Ridge Leadership Computing Facility (OLCF) deployed *Kafka* on top of *OpenShift Origin.*
  - Allows elastic scale out of Kafka nodes
  - With bigger system coming online, additional Kafka nodes will be deployed as necessary to handle increased volume of data
  - Conversely, additional Kafka nodes will be deployed to meet the increased demand from consumers

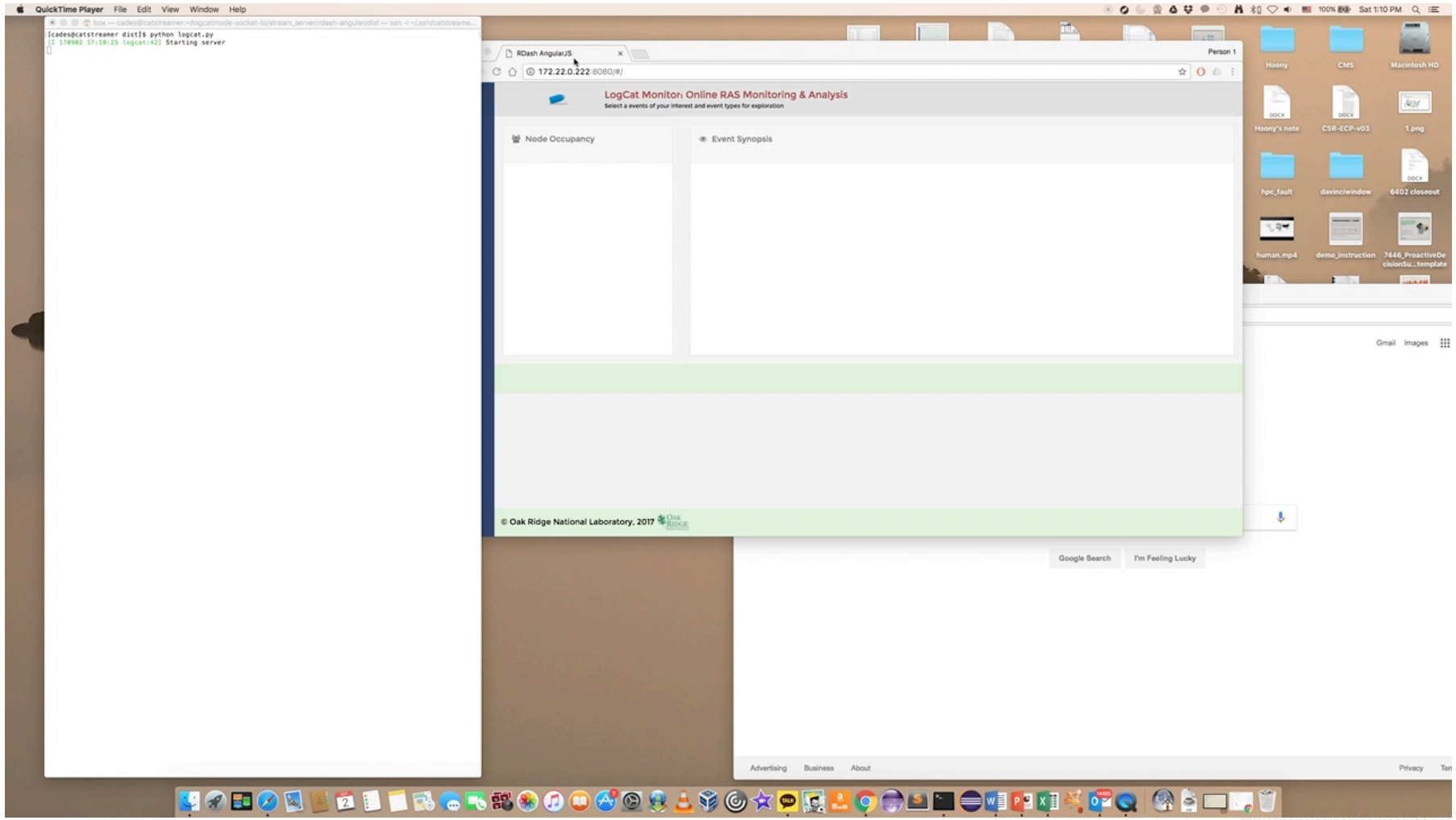




Online Monitoring & Analytics



OAK RIDGE National Laboratory



OAK RIDGE National Laboratory

Conclusion

- Big Data approach to designing and implementing HPC log analytics framework for large pool of users is presented.
- A-Z efforts that include:
  - Scalable data model for efficient retrieval from various perspectives.
  - Tightly coupled data and compute model for analytics
  - Frontend user interface for browsing and visualizing events with basic statistics computational capabilities.
  - On going online streaming processing for real time monitoring
- As exascale system coming available, the framework will play important role in understanding system health, status, and utilization.

OAK RIDGE National Laboratory

Acknowledgement

This research used the resources of the Compute and Data Environment for Science (CADES), and Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at the Oak Ridge National Laboratory, which is managed by UT Battelle, LLC for the U.S. Department of Energy, under the contract No. DEAC05-00OR22725.

OAK RIDGE National Laboratory

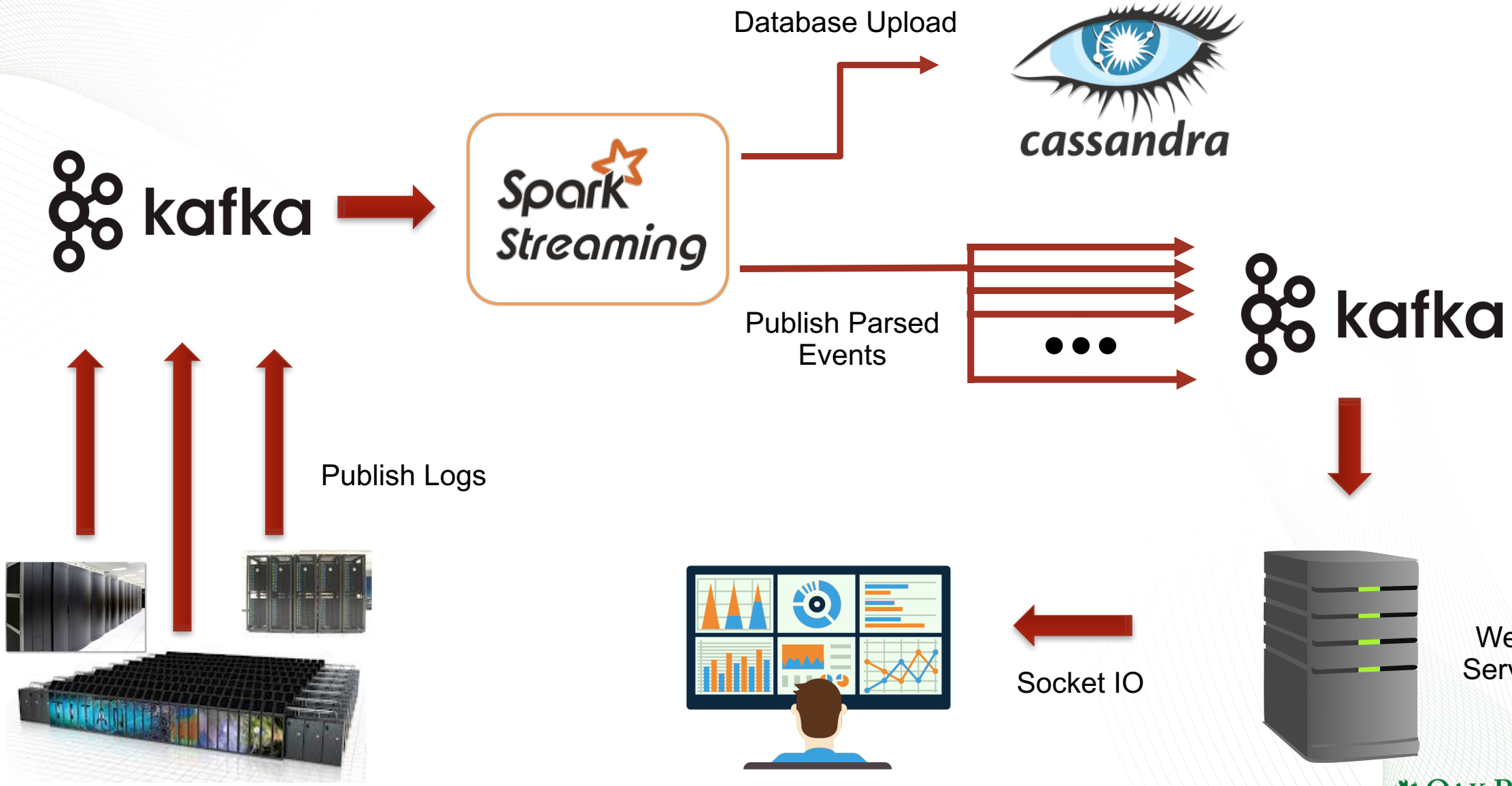
Questions?

Byung Hoon Park
parkbh@ornl.gov



OAK RIDGE National Laboratory

Online Monitoring & Analytics



OAK RIDGE National Laboratory

Relational and Log Data



- RDBMS needs to ensure ACID properties
  - ACID compliancy protects the integrity of your database by ensuring transactional interactions with the database.
  - Suitable for structured data that do not change and do not grow massively.
- HPC Log data processing does NOT require *ACID properties*
  - Does not need to update old data with transactional protection
- RDBMS for HPC Log data
  - Not flexible
  - Not scale

OAK RIDGE National Laboratory