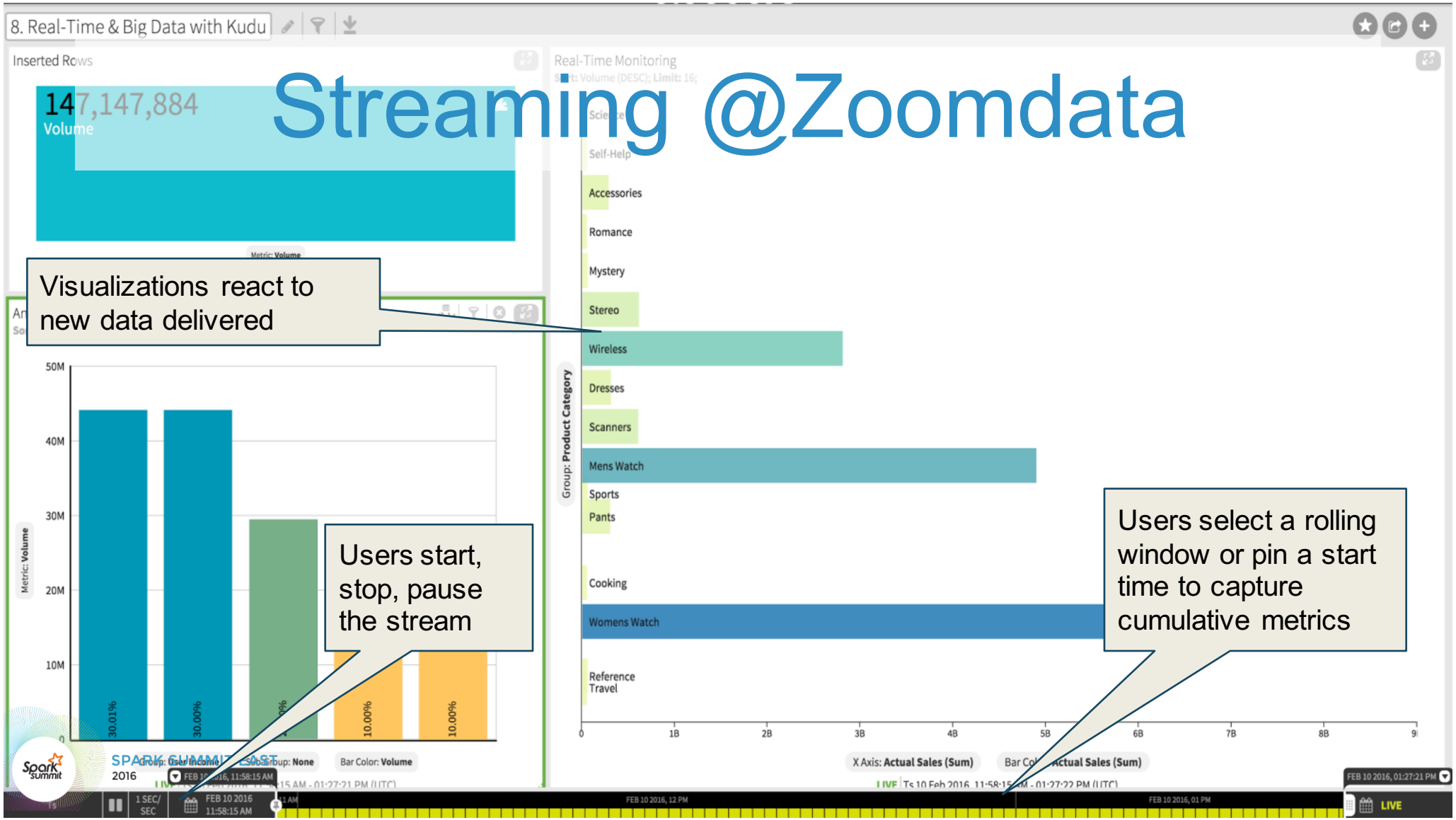
INTERACTIVE VISUALIZATION OF STREAMING DATA POWERED BY SPARK

SPARK SUMMIT EAST
DATA SCIENCE AND ENGINEERING AT SCALE
FEBRUARY 16-18, 2016 NEW YORK CITY


8. Real-Time & Big Data with Kudu
Inserted Rows
147,147,884
Volume
Real-Time Monitoring
Streaming @Zoomdata
Visualizations react to new data delivered
Users start, stop, pause the stream
Users select a rolling window or pin a start time to capture cumulative metrics
Self-Help
Accessories
Romance
Mystery
Stereo
Wireless
Dresses
Scanners
Mens Watch
Sports
Pants
Cooking
Womens Watch
Reference
Travel
Group: Product Category
Metric: Volume
50M
40M
30M
20M
10M
30.01%
30.00%
10.00%
10.00%
X Axis: Actual Sales (Sum)
Bar Color: Volume
1 SEC/ SEC
FEB 10 2016 11:58:15 AM
FEB 10 2016, 12 PM
FEB 10 2016, 01 PM
FEB 10 2016, 01:27:21 PM
LIVE

Drivers for Streaming Data

Data Freshness

Time to Analytic

Business Context

Challenges

- Time
- Frequency
- Retention
- Synchronization
- Order
- Updates

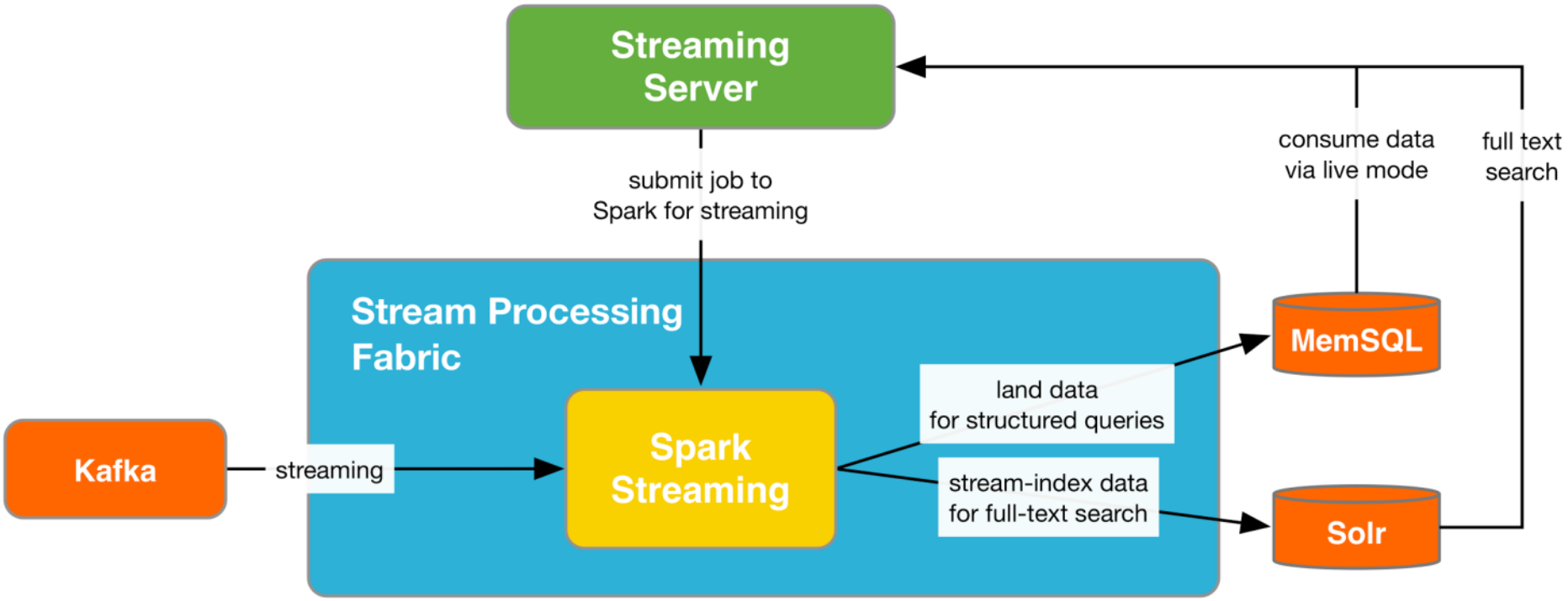
Addressing streaming @Zoomdata

	Historical	Revised
Receive Data	JMS	Kafka
Manipulate Stream	Single JVM in Memory	Spark Streaming
Hold Data in Buffer	MongoDB	Pluggable
Interact with Data	Custom Code	Pluggable

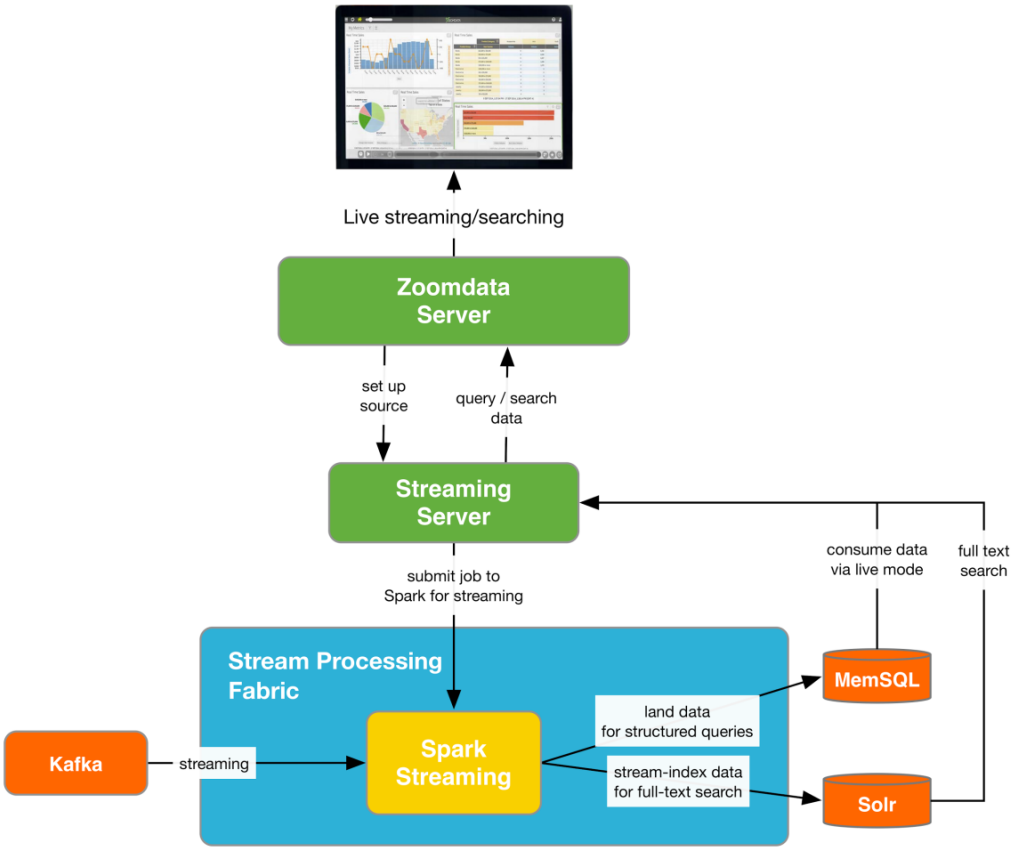
Technology Cast

- The Stream - Kafka, Kinesis, JMS
- Processing Fabric - Spark Streaming
- Landing Area - MemSQL, Solr, Kudu, Others

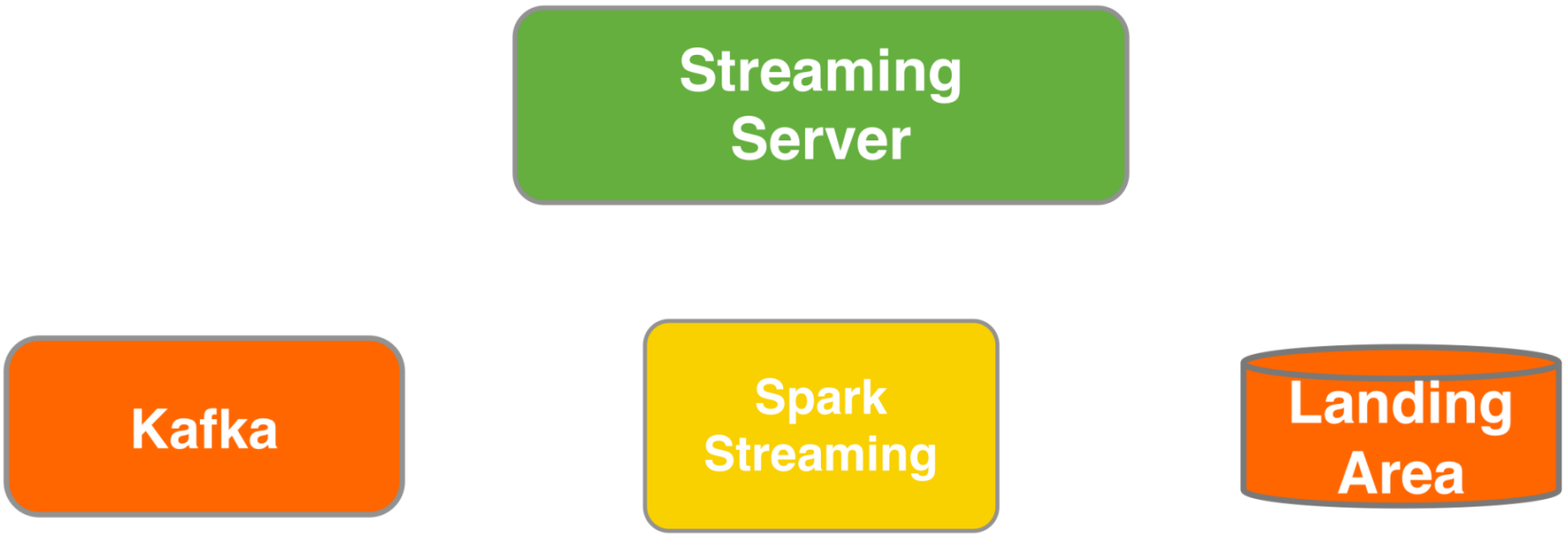
How it looks


Streaming Server
submit job to Spark for streaming
consume data via live mode
full text search
Stream Processing Fabric
Kafka
streaming
Spark Streaming
land data for structured queries
stream-index data for full-text search
MemSQL
Solr

With the rest of the app

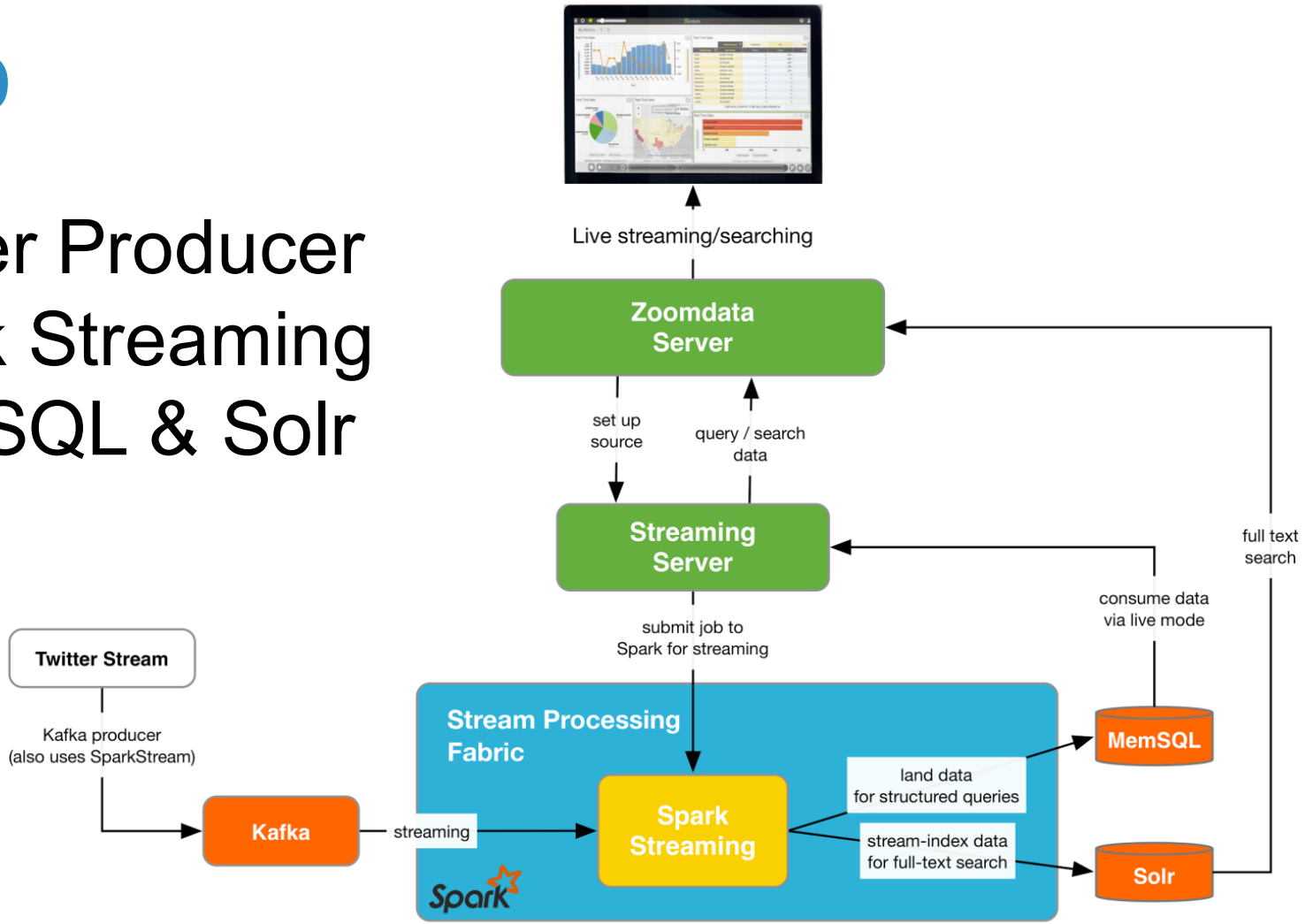


Scale Out



Demo

- Twitter Producer
- Spark Streaming
- MemSQL & Solr Sinks



Benefits

- Contextual Expressiveness with Streaming Data
- Independent scalability (scale-up, scale-around)
- Expressiveness powered by Spark -- using Windowing (dataframe API with stream)
- DR COOP, other Data management concerns

Future Work

- Cross stream synchronization & fusion
- On-demand scale out and resource management via Mesos
- Schema evolution
- More extensible landing strategies

Thanks