From Spark to Ignition:
Fueling Your Business on Real-Time Analytics
Eric Frenkiel, MemSQL CEO
June 29, 2015 • San Francisco, CA
What’s in Store For This Presentation?

1. MemSQL: A real-time database for transactions and analytics
2. Spark Use Cases
3. Example: Geospatial Enhancements
MemSQL Story

The real-time database for transactions and analytics
MemSQL at a Glance

- Experienced leadership from Facebook, SQL Server, Oracle, Fusion-io
- In-Memory, distributed, relational database
- Solving the Enterprise Architecture Gap
- Horizontal scale-out with modern database innovation
- $50 million in funding
Four Ways Your DBMS is Holding You Back

- ETL (Extract, Transform, Load)
- Analytic Latency
- Synchronization
- Copies of data

Source: Gartner Hybrid/Transactional/Analytical Processing Will Foster Opportunities for Dramatic Business Innovation
The Real-Time Database for Transactions and Analytics
The Real-Time Database for Transactions and Analytics

Data Loading and Queries -> Aggregator Nodes

Availability Group 1
Availability Group 2

memsql Cluster

Transactions

Analytics

Highest Value
Hot Data

High Value
Cold Data
Gartner Identifies Emerging Category: HTAP (Hybrid Transactional/Analytical Processing)

“HTAP will enable business leaders to perform...much more advanced and sophisticated real-time analysis of their business data than with traditional architectures.”

Download at: memsql.com/gartner
Simple

- Standard SQL
- Transactions and analytics in one database
- Behind the firewall or on the cloud
- Flexible integrations (Hadoop, Spark, SQL)
Fast

- Extremely low-latency queries
- Massive parallel transaction capacity
- Lock-free, shared-nothing architecture
Scalable

- Scales out on cloud and commodity hardware
- Deploys to thousands of machines
- True linear scaling
MemSQL Product Ecosystem

- Dashboards
- In-Memory Applications
  - Transactions and Analytics
- Advanced Analytics
- Streaming
- Wire-protocol compatibility
- Databases and Data Warehouses
  - ODBC, JDBC, .NET Connectors
- MemSQL Loader
- MemSQL
- Amazon S3
- Hadoop
- Streaming
- Azure
- Amazon Web Services
- Mesosphere
Spark Use Cases
Spark Data Processing Framework

Intuitive, concise, and expressive operations needed for analytics

Spark SQL
Spark Streaming
Mlib (machine learning)
GraphX (graph)

Apache Spark
Understanding MemSQL and Spark

Cluster-wide Parallelization  |  Bi-Directional
Spark with MemSQL

MemSQL Spark Connector enables the real-time trinity

Message Queue: Kafka
Transformation: Spark
Data Serving: MemSQL

Programming libraries
Persistence
Application platform

End-to-End Data Pipeline Under One Second
MemSQL and Spark Use Cases

- Operationalize models built in Spark
- Stream and event processing
- Live dashboards and automated reports
- Extend MemSQL analytics
Operationalize Models Built in Spark

- Process in Spark, persist to MemSQL
- Go to production and iterate faster
Stream and Event Processing

- Structure event data on the fly
- Pass to MemSQL for persistent, queryable format
Real-Time Analytics at Pinterest

- Higher performance event logging
- Reliable log transport and storage
- Faster query execution on real-time data

50,000 pins/sec
Live Dashboards and Automated Reports

- Serve live dashboards from MemSQL
- Run custom reports on live data with Spark
Extend MemSQL Analytics

- The freshest data for analysis in Spark
- Load from MemSQL to Spark and write results on return

Interactive Analytics, Machine Learning

Access to Live Production Data

Real-time Replica

Applications, Data Streams
MemCity

- Capturing energy consumption data from 1.4 million households
- 8 devices per household
- 186,000 events per minute
- AWS hardware costs at $2.35 per hour
Kafka 322.0  
Spark 330.0  
MemSQL 338.0  

27,100,500,917  
Megawatt-hours

Consumption by device:
- Laundry & Water
- Kitchen Appliances
- misc
- Entertainment
- HVAC
- Lighting
- Small devices

141,041,035  
Database records
Geospatial Challenge

- Commercial applications *now geo-enabled*
- Location is everywhere
- Lots of insight possible
- Traditionally geo is processed separately
- Real need for integrated geospatial at scale
MemSQL Geospatial

- Points, Lines, and Polygons
- Topological filters
- Measurement functions
MemSQL Geospatial

- BILLIONS of objects
- Sub-second latency
- Geo data is first-class citizen
- Geo + Simplicity + Speed + Scale
Real-Time Geospatial Location Intelligence

- Sample from 170 million taxi trips
- Real-time ingest
- Concurrent queries in fractions of a second
- Unlimited number of geographic views
- Simple queries while simultaneously ingesting data
SELECT SUM(passenger_count) as result
FROM trips
WHERE GEOGRAPHY_INTERSECTS(pickup_loc)
OR GEOGRAPHY_INTERSECTS(dropoff_loc)

RUN QUERY
A database so scalable that everyone can use it.

UNLIMITED scale and capacity

Free FOREVER
Thank You!

Visit the MemSQL Booth #4

MemCity Showcase  Games  Giveaways