JBoss
Community

Big Data @ Red Hat

Research and Development Directions

Jonathan Halliday
jonathan.halliday@redhat.com

August 2013

Big Data @ Red Hat

- Existing work
  - GlusterFS distributed file system (posix)
  - Infinispan data grid (jsr-107)
  - Modeshape distributed doc store (JCR)
  - Hibernate OGM & Search
  - Drools Fusion (CEP)
  - Teiid database federation

Big Data Analytics @ Red Hat

- Analytics capability
  - Complement existing efforts
  - Layer on existing nosql storage
  - 'real time' focus
  - Product and Service
  - Research and Development

Recording and Analysis

- Numeric Input Data
  - Fixed interval sampling
  - Event capture
- Applications
  - System monitoring
  - User activity tracking

Recording and Analysis

- Volume and Velocity
  - Shorter sampling intervals
  - Greater event resolution
- Visualization
  - Charts and dashboards
  - Real-time mining and updates

Recording and Analysis

- Triggers and Alerts
  - Starts to look like stream processing / CEP
  - Hybrid stream / historic data use
- Predictive Analytics
  - Demand estimation, capacity planning
  - Cycles and noise
  - Context awareness vs. pure maths

Implementation

- Map/Reduce is too slow
  - Although not as bad as it was
- Sacrifice generality for speed
- Query templates
- Query oriented storage
  - Layout to minimise disk I/O
  - denormalization

Implementation

- Cassandra
  - Distributed column family database
  - Dynamo distribution, Big Table datamodel
  - Great write scaling, including counters
  - CQL3
  - 2.0: adds CAS, triggers

Perspicuus

- SQL/CQL like DSL for data cubes

```
STORE SUM(<numeric_property>)
FROM <event_class> INTO <table>
GIVEN <some_property>
GROUP BY <other_property>

SELECT FROM <table>
WHERE <some_property>='x'
```

Academic Connections

- MSc coursework
  - CSC8101: Big Data Analytics
  - CSC8104: Enterprise Middleware
- Sponsored PhDs
  - 3 in progress
- EU research projects
  - Cloud-TM, LEADS

Related Research

- Rebecca Simmonds – CS PhD
  - In progress, 2/3 done
  - Identify design patterns for storage and query
  - Social Media (twitter) proof of concept application
  - Perhaps further streaming/historic join work

Related Research

- Rui Vieira – CS MSc, Maths PhD
  - Industrial placement on distributed top-k over cassandra
  - Doctoral work initially on predictive analytics
  - Multi-disciplinary nature of data science

Related Research

- Cloud-TM
  - EU funded work on “Self-Optimizing Distributed Transactional Memory middleware”
  - Feature enhancements for Infinispan
    - Atomic broadcast
    - non-blocking state transfer

Related Research

- LEADS: Large-scale Elastic Architecture for Data as a Service
- Federated micro-clouds
- Continuous and snapshot queries
- Historical and streaming data
- Infinispan, OpenShift

JBoss
Community