Hadoop as a Service

How to build & operate a multi-tenant enterprise data lake supporting operational & streaming analytics

11:15–11:55 Capital Suite 10/11

Phill Radley
Chief Data Architect - BT

These slides @
conferences.oreilly.com/strata/strata-eu/public/schedule/detail/57589
Whose in the room?
BT’s big data journey
BT Research & Innovation

Apache Hadoop
Experimental Cluster ~ 2008

2010 Community Edition
Amazon Web Services
Twitter

Cluster 1 Daily Job Count

Jun 2012 May 2016

Feb 2014 – map/reduce + HDFS

Oct 2016 – Spark + Impala + Kafka ~ 2.5x

#StrataData

Production
Connecting legacy applications/data to HaaS

104 Systems connected across 7 business units
Typical usecase 1... **nuisance calls**

- **Machine Learning Model**
  - Calculates Nuisance Score

- **Call Features**
  - Duration
  - Valid CLI
  - Who hangs up

- **Non-linear transform**

- **Nuisance Score**

- **Black List of Nuisance Call Centres**
  - (customer opt in)

- **Divert call to Junk voicemail box**

---

JOIN CHRISTINE, FIGHT NUISANCE CALLS

#StrataData
Typical use case 2... Broadband Speed

Predict when engineer visits will improve speed (each dot is an engineer visit)
Current Technology Stack

Hadoop as a Service (big data platform)

Data Warehouses & Marts

Legacy Applications

OLTP APPs

CRM/Billing/….

Hadoop Distributed File System

Spark-submit (create context)

Start Flink cluster

Map Reduce batch job

Fair Share Scheduler

Hive db

Hive db

Hive db

Hive db

Hive db

Hive db

Hive db

YARN

Guaranteed ££

PROD 80%

NON-PROD 20%

Dedicated / Multi-Tenant

Encryption at Rest

HttpFS

Parquet

Apache Flume

Apache Hadoop

Cloudera Distribution

httpFS

Active Directory

10GB LAN

Cisco

Hewlett Packard Enterprise

Oracle Enterprise Linux

ORACLE

BUSINESS INTELLIGENCE

Shiny

Datameer

Data Policy Engine

ORACLE

GOLDEN GATE

Edge Clients

Data Policy Engine

Legacy Applications

OLTP APPs

CRM/Billing/….

Start Flink cluster

Spark-submit (create context)

Map Reduce batch job

Guaranteed ££

PROD 80%

Non-PROD 20%

Dedicated / Multi-Tenant

Encryption at Rest

HttpFS

Parquet

Apache Flume

Apache Hadoop

Cloudera Distribution

httpFS

Active Directory

10GB LAN

Cisco

Hewlett Packard Enterprise

Oracle Enterprise Linux

ORACLE

BUSINESS INTELLIGENCE

Shiny

Datameer

Data Policy Engine

ORACLE

GOLDEN GATE

Edge Clients

Legacy Applications

OLTP APPs

CRM/Billing/….
Cluster Utilisation: Weekly

<table>
<thead>
<tr>
<th>Date</th>
<th>Containers Running</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon 13</td>
<td>350</td>
</tr>
<tr>
<td>Tue 14</td>
<td>300</td>
</tr>
<tr>
<td>Wed 15</td>
<td>250</td>
</tr>
<tr>
<td>Thu 16</td>
<td>200</td>
</tr>
<tr>
<td>Fri 17</td>
<td>150</td>
</tr>
<tr>
<td>Sat 18</td>
<td>100</td>
</tr>
<tr>
<td>Sun 19</td>
<td>50</td>
</tr>
</tbody>
</table>

Max
Mean
Daily "Quiet" time
3-6am
Min

#StrataData
Cluster Utilisation: Daily

Total Containers Running

Max
Mean
Min

Cluster Utilisation:
- Quiet
- Busy
- Busiest

#StrataData
Cluster Utilisation: “Busy Hour”

Tenants Running 😊

Tenants Waiting 😞

1 hour preceding July 1, 2016, 9:56 AM BST

cleared by 10:25
- “One size fits all”
- daily provision run
- Terms & Conditions

Tenant On-boarding

End User Groups (self-service)

Application Teams

HAAS

#StrataData
Hybrid Architecture

Legacy systems
Adoption / On-boarding Patterns

IT led projects

Business led projects

Governed Self-service
Data & Analytics

Business Question or decision?
Hadoop as a Service

Questions?

Meet the Expert with Phillip Radley (BT)
12:05–12:45 Wednesday
Capital Hall (N24) - O'Reilly booth - Table B

Rating & slides @
conferences.oreilly.com/strata/strata-eu/public/schedule/detail/57589
phillip.radley@bt.com
Hadoop as a Service, how to build & operate an enterprise data lake supporting operational & streaming analytics

- Migrating from MapReduce to YARN, Spark, and Impala
- Migrating a live cluster to a new hardware stack
- Adding HDFS encryption
- Adding in second site for disaster recovery (adopting Airbnb’s gold/silver cluster approach)
- Evolving the architecture to support streaming applications (using Kafka, Spark, and Flink)