Enterprise Architecture Patterns For Big Data

Phil Radley,
Chief Data Architect
20 / October / 2016
What I’m going to talk about....

- The organisation of BT and its IT.
- Early stages of big data in the industry & BT
- The BT production big data platform (HaaS)
- Sample Adoption Patterns
  - Data Archive, DW Extension, Re-Platform old batch apps, Self-Service Analytics
- Example Use Cases
  - Copper Line Performance Model – Broadband Speed Prediction
  - Nuisance Call Prediction
- Governance
- Where next
BT Group Structure  1/Apr/2016

Customers

Openreach  Wholesale & Ventures  Global Services  Business and Public Sector  Consumer  EE

Technology, Service and Operations

Research & Innovation  Chief Architects Office  Enterprise Architecture

Group Functions

HR, Finance, Corporate Affairs, Brand, Customer Experience, Strategy, Portfolio and Policy, Legal, Governance and Compliance

Revenue

£23.7Bn

Openreach  BT Global Services  BT Wholesale  BT Consumer  BT Business

~ 90K FTE in 61 countries, serving 180 countries
Legacy Systems Architecture in each BT Business Unit

- Hundreds of systems in each business unit grouped into 3 operational areas (CRM/Service Mgt/Network Mgt)
- Data Warehouse per business unit
- Client – Server applications running on servers in BT Data Centres (~ 35K hosts)
- Mainframe applications (in Openreach)
- Total Storage ~ 25PB
- Lots of event / time series data
  - Network Alarms & Telemetry
  - Netflow Traffic Events, Security events
  - Call Detail Records, web clicks,
  - mobile handset data (GPS, Apps, browsing..)
- Business Unit CIOs manage IT investment roadmap, each business unit deploys a “stack release” quarterly
Research & Innovation (R&I)

- R&I is a unit within IT (TSO) located in Adastral Park Campus
- ~10 practices progressing research topics annually agreed with business units
- External innovation team (based in Silicon Valley + Boston(MIT))
- Big Data & Customer Experience (incl/ social media) Practice
  - established ~ 2008
  - 20 people
  - First Hadoop Clusters on AWS ~ 2009
  - Migrated to on Premise 2011 (Cluster 1)

Research Cluster 1 closed after 4 years / 250K map/reduce jobs
The Long View of Big Data

Data “Bigness” = \( f(\text{Volume, Velocity, Variety}) \)

- **Mainframe (1st Platform)**
  - COBOL/ISAM/IDMS
  - Linked Record sets
  - Proprietary, Monolithic
  - Batch, Interactive

- **1960**

- **1990**

- **Y2K**

- **1990**

- **2006**

- **2009**

- **2014**

- **2016**

- **Open! 3GL, 4GL PC & Servers on premise**

- **Relational**

- **Clusters, Data hub, pipelines**

- **Mobile Social Big Data Cloud**

- **BT**

- **First Research cluster**

- **Production Cluster**

- **HAAS = Hadoop as a Service**

**Battles for Dominance and Survival — on the 3rd Platform**

**Idc Predictions 2014**
Early Big Data in BT - 2011-14

- “Head Office” gave R&I big data practice “fuzzy business problems” to analyse
- Data Science team manually assembled relevant data sets and worked on them to produce correlations, joins and predictive models
- Being outside core IT simultaneously constrains and liberates possibilities
- By 2013 the business units were starting to rely on R&I Hadoop and a production capability was needed
Feb 2014 Production Launch “Hadoop as a Service”

- Following a presentation to the TSO Leadership team Dec 2013 an initial investment in a production cluster was agreed backed by a plan to launch in Feb 2014
- 60 nodes optimised for Hadoop map/reduce deployed in BT Data Centre in Sheffield (6TB local disks, 1:1 core:spindle ratio, 8GB for JVM per map/reduce slot
- Existing linux 3rd line team tasked with running basic (Min. Viable Product) Hadoop Cluster as a shared service platform
What is a “Shared Service Production Big Data Platform”

• User Authentication & Authorisation handled by Active Directory (integrated with Kerberos)
• Hadoop Distributed File Store
• Map/Reduce – Parallel batch processing framework + command utilities (Hive, PIG, sqoop, oozie)
• User Access via telnet/linux Command Line + Browser Based front end (HUE)
• Data Transfer (batch files via httpFS, Oracle Tables via Sqoop (now Golden Gate), flume for telemetry)
• Single Page Intranet Order Form
• Standard IT Helpdesk (similar to infrastructure services linux, Oracle, WAN)
• Two categories of platform users
  1. Developers/Testers/support working on applications using Hadoop to store or process data
  2. End users consuming data, 3 broad groups of end users :-
     i. Handful of Data Scientists using tools like R-Studio
     ii. Tens of business analysts using SQL + Hive (Data warehouse sandbox users)
     iii. Tens of simple users working on specific business problems/Questions (using Datameer or in a team with (i) & (ii)
Multi-Tenant Hadoop as a Service

HAAS Platform

Orders form (SharePoint)

HAAS Platform

Analytics Review Board

HAAS Platform

Tennant “Project Owner”

“HAAS AP 00307_12126 Is ready for you to use”

User admin

User Access

Active Directory

Hadoop Cluster B (Openreach only)

Standard User Admin Process

Script

Hadoop Cluster A

Job queue

HDFS

HUE

Impala

Datameer

BI Server

Oralce DB

Flume

http FS

Squop

APP extends footprint in HaaS

Systems Access

Production

Cloudera manager

Cloudera

Kerberos

Create Security Group

HAAS Platform

Create Hadoop Features
Service identifiers (link to Architectural Repository)

HAASA AP 00307_12126

Prefix Identifies Hadoop Groups In Active Directory

Cluster (A or B) Service Type
Application(Dev/Test/Prod) End User

Service Instance No. (order number)

Suffix
“Application ID” Link to legacy application (you do have an Application register?) OR..ARB board approval number R_0030

Service ID = Active Directory Group Name
A Modern Enterprise Data Architecture (V1.0)

1. Event Ingestion from Networks/IT/Web servers Collection with flume agents landing in HDFS files

- CRM
  - Web/APP Server
  - RDBMS
  - Golden Gate
  - Map
  - Reduce code

- ERP
  - Web/APP Server
  - RDBMS
  - sqoop
  - Map
  - Reduce code

- DW
  - Web/APP Server
  - RDBMS
  - sqoop
  - Map
  - Reduce code

2. DB Table transfer using sqoop (map/reduce) jobs, landing in HDFS files

- CDC snapshot
- HDFS snapshot

- Active Directory

- Wrangling & Discovery
  - Data Science
  - Datameer, HUE...
  - (HDFS FILE ACCESS)

- BI Tools
  - Tableau, Zoomdata...
  - (HIVE TABLE ACCESS)

Insert Update Delete into RDBMS
“Active Data Archive” Pattern

- Simple starter pattern for help application designers solve a common problem (long term archiving & managing data retention)
- Brings useful data sets to the cluster
- Can be used to provide a central archive of a particular data set, e.g. VAT Transaction archive (saves clogging data warehouses)

Raw tables (OLTP schema) sqoop’d onto HDFS and transformed into reporting schema with pig/hive. Daa exposed to range of tools by ODBC.

Auditors + analysts

BI Tools
Tableau, OBI...
(HIVE TABLE ACCESS)
Data Warehouse Extension

- DW Sandbox offload to Hadoop, esp. heavy ad-hoc DW users
- ETL offload from ETL/ELT servers to Hadoop (Data Reception Area)
  - Faster integration of new sources (schema on read, Table oriented)
- Archive Data from DW

Power user Sandbox offload to Hadoop

BI Tools
Tableau, OBI...

Analysts
Re-Platform old Batch Applications
Customer Data MDM Hub (CMF)

MDM hub for Business Customer Master file (CMF)
- 10 years old and needed re-platforming (2014)
- 12 Source systems with local customer table
- D&B Legal Entities used as Reference Data
- Existing modules ported to Hadoop/Hive

Benefits
- Business able to multiple runs in a day
  Hadoop 15x faster
- Cost saving over standalone re-platform
- Data volumes increased 3x (multiple i/p files)
- Adding new sources is quicker (schema on read)
- Data available for Self-Service Teams (DQ/Data Science)

- Using this “ETL Offload” Pattern the Master Address data is being converted to a hybrid hadoop application, Data migration

✓ Hybrid apps are a low risk entry route to big data
Copper Line Testing & Performance done as **Self Service Analytics**

- Data Warehouse
- NETWORK INVENTORY
- ADDRESS
- LINE TESTS
- DSL LINE SPEED

Calculate 25 Million line lengths

Join with line tests & DSL line speed data

Anomalous Test Results

Reporting Server (Shiny)

Self-Service Data Analysis Team

Exchange Engineers

Datameer
### Data Wrangling

- **Distribution**
- **Row Count**
- **Unique Values**
- **Min**
- **Max**
- **Mean**

#### Data Profile

- **Category**
- **Old**
- **Remarks**
- **bit_ownership**
- **bit_ownership**
- **Shape**
- **bit_plant**
- **bit_objec**
- **exchange**

<table>
<thead>
<tr>
<th>Category</th>
<th>Old</th>
<th>Remarks</th>
<th>bit_ownership</th>
<th>bit_ownership</th>
<th>Shape</th>
<th>bit_plant</th>
<th>bit_objec</th>
<th>exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
<tr>
<td>DP</td>
<td>1,000,000</td>
<td>[DP TYPE] [PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
<td>[PR]</td>
</tr>
</tbody>
</table>

**Only needs Excel style skills + knowledge of business data**
Broadband Speed Prediction
(data from old systems loaded onto HaaS & analysed in new ways)

Computed Daily Based on
5M daily line tests

Proof the model works
(blue dots = wasted truck roll
= £M in savings)
Data Visualisation for Exchange Engineers (R-Shiny Server)
Nuisance Call Scoring Model (BT Saturn)

Provision, Load & Model with 2BN CDRs in ~ 2 weeks
HaaS Platform Growth & adoption

- Sept 2016
  - 82 Systems “connected” to the data hub
- 400 services provisioned

Graphs showing growth from Feb 2014 to Sept 2016.
### Governing Use of the Platform (Project/System On boarding)

**IF** `{project name}` **is following standard IT Operating Model*:**

<table>
<thead>
<tr>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document use of HaaS in the design, use the approved patterns</td>
</tr>
<tr>
<td>Order HaaS from Data Centre team</td>
</tr>
</tbody>
</table>

**ELSE:**

<table>
<thead>
<tr>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register <code>{project name}</code> with Analytics Review Board</td>
</tr>
<tr>
<td>IF Approved (Order HaaS from Data Centre team)</td>
</tr>
</tbody>
</table>

---

Note: This is just governing use of the platform, Data Governance and Compliance is a whole separate topic! Eu GDPR, PCI-DSS, MDM etc. etc.
Three existing business applications (CRM, Orders, Faults) extended into HaaS.
What’s coming next? Event Streaming for Time Series Data

Data Streams replaces batch mode

“The Dataflow Model”
proceedings of the VLDB Endowment, vol. 8 (2015), pp. 1792-1803

Most analytics done on time series data (events as they arrive)
...web clicks, call detail records, IOT alarms & usage, vehicle telemetry, mobile GPS..

Flink
beam
kafka
KUDU
Enterprise Architecture Patterns For Big Data

Q & A

Phill Radley
Chief Data Architect
phillip.radley@bt.com