THE BANK OF ENGLAND’S BIG DATA JOURNEY
Introductions

BARRY WILLIS
- Worked as a developer for 13 years, before moving into Team Leading, for the most part in Data Analytics & Modelling
- Data Platform Lead for the Bank of England since 2016

NICK VAUGHAN
- Joined the Bank of England in 2012 as a BI Specialist
- Previous development, architecture and strategy roles in Telco, Retail, and Construction industries.
- Domain SME for Data Analytics & Modelling
What are we going to talk about?

- History of Bank of England’s Data Platform
- Our steps to delivering a Big Data Use Case
- Next steps in the journey
Part of the strategic review was to look at potential data strategies, with a recommendation initially delivered in 2015. Key messages were:

- Improve ability to share data across the Bank – reduce data silos / numbers of systems
- Improve analytical capabilities across the Bank – e.g. share tooling / sandboxes etc.
- Support genuine Big Data use cases
- Strategic plan covered these strategic data themes:
  - Management [Governance & Security]
  - Collaboration [Sharing of Data]
  - Standardisation [More robust processing]
  - Exploitation [Tooling for gaining data insight]
- A 3 year Data Programme began in 2016 to do something about this
Stage 1: The Appliance / Data Hub...
...and wider components

- **Virtualised Data Lake**
  - Central Hub for data
  - Production Workflows
  - Certified datasets
  - Tuned hive databases
  - Direct connection

- **ETL Tool**
  - Ingestion
  - Transformation
  - Tokenisation
  - Metadata Management (never implemented)
  - Data Lineage (never implemented)

- **BlueTalon**
  - Access management
  - Security
  - Dynamic data obfuscation
BIG DATA CASE STUDY
What was it?

How much data?
- The biggest we’d ever had! Daily load compares to around 3 months worth of data of our second biggest data analytics solution
- Around 80 daily files containing 30 million rows of derivative transactions. Total is approximately 20GB per day
- 3 years of historical data to be loaded in total (currently we have 9 months)

How is it being used?
- Users are aggregating data in multiple different ways, which can be extended to more unspecified questions to be answered in the future.
- We had no other solution in the Bank to enable users to easily analyse this amount of granular data – too big for traditional data warehouses.
Re-engineering: Low Level Design

Data Governance

Landing Zone
- Raw Zone
  - Structured Zone
    - Refined Zone
      - Consume Zone

TR State Data
- zip
- csv

External Reference Data
- zip
- csv

Spark
- orc

Label: Out of scope
In scope
Case Study: Outcome

BUSINESS
• End users are able to complete analytics in 5 minutes which previously took weeks
• Solution reduces opportunity for data preparation errors
• One use case delivered, 6 more planned and up to 30 in total

DATA
• Already seeing other business areas wanting to make use of the new data in the Lake.

TECHNICAL
• We’ve learnt huge amounts about Data Engineering
• The framework design is reusable across different data collections, large or small
• We’ve built the basis of a reusable Data Quality framework
NOW AND NEXT...
Continued Platform Challenges

- Our Appliance has been made end of life by supplier
- Disaster Recovery – Unable to use for critical systems
- Data Catalogue / Search and Metadata Management
- Environments tuned for function – ETL / Query segregation
- Sandboxing & End User Data Prep tooling
- Big data continually evolves – e.g. Hortonworks / Cloudera merger

- Operating models
- Culture & Education
What's Next: Data Hub 2