Role of Data Engineering in Data Science
Bhagya Reddy
Principal Data Engineering

Studied Masters in Mathematics

Before joining QB I worked with multinational companies like Wipro, IGATE, UBS, GE as in BI Engineer till 2012

I started working as Data Engineer from 2013 working in Specsavers and Carbon Credentials

I am experienced in BI data projects in Banking, Commercial, Energy and Sustainability

I joined QB in October 2016

After joining QB I expanded my domain experience to include Healthcare, Banking, Telecoms and Energy industries

I also started working alongside Data scientists and providing cutting edge solutions
Today’s agenda

- What is Data Engineering and how we define it?
- Our pain points a couple of years ago
- Feature Engineering
- Few client studies examples
# Data is everywhere and hugely increasing day by day

Data is transforming businesses …  
… to yield immense impact

<table>
<thead>
<tr>
<th>Company</th>
<th>Description</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google</strong></td>
<td>Analyzed 600+ variables to predict propensity to respond &amp; spend for small- and mid-size business customers</td>
<td>5.7x increase in customer ad spend</td>
</tr>
<tr>
<td><strong>Amazon</strong></td>
<td>Instituted a real-time, individualized customer recommendation engine powered by machine learning</td>
<td>25-35% sales driven by recommendations, $15B-$25B revenue derived from recommendations</td>
</tr>
<tr>
<td><strong>Boeing</strong></td>
<td>Developed and deployed real time analysis of sensor data to optimize ideal vehicle velocity at various inclines</td>
<td>17% decrease in fuel consumption, &gt;80% reduction in 787 assembly time</td>
</tr>
<tr>
<td><strong>UPS</strong></td>
<td>Implemented an automated, real-time advanced route optimization engine for delivery vans (including elimination of 90% of left turns)</td>
<td>30M decrease in miles travelled, 5000 trucks removed from the road</td>
</tr>
</tbody>
</table>

Instituted a real-time, individualized customer recommendation engine powered by machine learning

Source: Press search

---

**Data, rather than software, is the defensible barrier for many businesses**

– Andrew Ng
Our definition works like this...

Data Engineering
- Data landscape mapping
- Construction of pipelines
- Feature engineering

Data Science
- Feature engineering
- Model selection
- Model performance

Machine Learning Engineering
- Productionisation
- Performance optimisation
- Software engineering rigour
Digital projects require new roles spanning three core business functions.

Ensure future data requirements and processing are robust and complete.

Translate business needs into digital language (e.g., by defining data requirements).

Develop insights leveraging new sources of data and best-in-class algorithms.

Roles:
- Business-domain expert
- Product owner
- Functional expert
- Business leaders
- Data engineer
- Translator
- Software/UX developer
- Business leaders
- UX/UI designer
- Robotica & automation expert
- Agile coach
- Scrum master
- Data architect
Some of our main pain points a couple of years ago

😢 Each project was designed for the problem at hand

😢 Data Engineering was done in Scala or SQL

😢 Data Science was 60% R vs 40% Python

😢 Communicating the value of Data Engineering was / is hard

😢 Our projects were often ‘proof of concepts’ rather than full interventions

😢 Swapping people in and out of projects was a headache

😢 We found ourselves doing similar things differently over and over again
Data Engineers Core skills

Core Skills

• Client facing:
  – Interact closely with clients at a technical level to understand the data and prioritise data sources based on value
  – Navigate and agree Information Security controls for secure data exchange and processing

• Technical experts:
  – Expert in joining very disparate data sets (structured and unstructured) of varying quality to bring unique perspectives (which in itself can add huge value)
  – Do the data ‘heavy lifting’ to free up the Data Scientists to focus on modelling
  – Build reusable data pipelines that can quickly be converted

• Eye for efficiencies:
  – Develop domain data models to speed up data acquisition and feature engineering for further industry use cases

Expertise in programing and big data tech stack (Examples)
Different staffing models are available to deliver different parts of the process

You can staff Data Engineers for all or parts of the process depending on your needs, e.g.¹:

<table>
<thead>
<tr>
<th>Identify Data</th>
<th>Securely Extract and Load data</th>
<th>Clean, Join, and Validate Data</th>
<th>Feature Engineering</th>
<th>Modelling / Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE for scoping</td>
<td>DE for ingestion</td>
<td>DE for ingestion through to Feature Engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹ Some examples – not exhaustive. We can help recommend the best working model to meet your requirements
The Data Architecture needs to scale with a well-governed, reusable, and auditable approach and align to data management.

1. Establishes a common domain-modelled primary layer serving multiple use cases at scale, which has a life beyond models.
2. Creates reusable feature stores structured by use case or domain for simplicity and traceability.
3. Enables effective auto-scaling, and clear model versioning and monitoring for safe live models.

**Diagram Description:**
- **Ingestion Layer:** Raw data from both internal and external sources is ingested.
- **Primary Layer:** Duplicated and complex data is transformed into cleansed, joined, transformed data.
- **Feature Layer:** Multiple feature stores with inconsistent data are created, enabling effective auto-scaling.
- **Modeling Layer:** Models are built, trained, and evaluated on consistent data.
- **Deployment and Serving Layer:** Application developers can access feature stores through an API, with monitoring services ensuring safe live models.

**Key Points:**
- **Reusable data asset**
- **Data Science**
- **DevOps Engineering**

**Technical Details:**
- **RAW TEXT END**
Let’s dive a bit deeper into DE responsibilities

- Subject matter expert / data owner interviews
- Understanding data landscapes
- DQ report & profiling
- Data diagrams & Data dictionary
- Ingesting data

- Cleaning up data not designed for analytics
- Collaborate as part of the hypotheses definition process
- Build the data architecture & engineer features
- Build the model input table
- Build end to end pipeline for handover
What the Data Engineering team brings to the Feature Engineering process

Business SME’s help us identify data sources with which can test our hypotheses.

- We map the data sources against potential hypotheses to understand ‘the art of the possible’
- We use Studio to track hypotheses and data sources and the features
- Writing Spark / Python / SQL pipelines
Tips for data engineering

- Use a template to structure your project repository.
- Use Jupyter notebooks for exploration and storytelling, pipelines for production code.
- Use reproducible environments.
- Version control your code. Data is immutable.
- Don’t repeat logic, write utils that could be assets across projects.
- Abstract the business definitions in configuration files.
QB exploit data, analytics and design to help our clients be the best they can be

We were born and proven in Formula One, where the smallest margins are the difference between winning and losing and data has emerged as a fundamental element of competitive advantage.
### Few example case studies

These client success stories would not have been possible without Data Engineering.

<table>
<thead>
<tr>
<th>Industrial</th>
<th>Finance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predictive Maintenance on Mining Trucks</strong></td>
<td><strong>Insolvency Prediction</strong></td>
</tr>
<tr>
<td>Global mining company wants to reduce the inoperable time of their equipment and machinery, due to repair requirements</td>
<td>A government body wanted to fundamentally change how it approached tax debt collection to reduce the significant tax debt write-off due to insolvency</td>
</tr>
<tr>
<td>Extracted data from client systems and decoded into human-readable information</td>
<td>Identified and consolidated siloed data</td>
</tr>
<tr>
<td>Developed a parallel compute solution (Spark) to process and aggregate the data in hours instead of days/timing out using machine learning methods</td>
<td>Navigated a large government organization to locate disparate data sources and subject matter experts to develop a ‘single view of a company’ that had never been achieved before by the client</td>
</tr>
<tr>
<td>Optimised feature engineering to function across large data sets, enabling data science models to predict 85% of truck exhaust failure with a lead time of 6 days</td>
<td>Deployed a live tool</td>
</tr>
<tr>
<td>Enabler to model impact</td>
<td>Built and deployed the single view application to 6,000 collectors. The client stated ‘It’s a game changer – it’s revolutionary for us’</td>
</tr>
<tr>
<td><strong>Engineering Performance</strong></td>
<td><strong>Client capability Building</strong></td>
</tr>
<tr>
<td>A global engineering consultancy wanted to understand the drivers of performance across their three business lines, aiming to achieve a strategic target of ‘10% profit on income’ on a sustainable basis</td>
<td>Educated the client on data matching and validation: the DE approach was presented to directors and analytics practitioners across the organization, training the client to adopt a better process</td>
</tr>
<tr>
<td>Aggregated data covering 12,000 projects to evaluate the variance of project profitability and cost across the three business divisions</td>
<td><strong>Pharma Salesforce Effectiveness</strong></td>
</tr>
<tr>
<td>Analyzed to quantify the value of projects by fee size</td>
<td>A global pharmaceutical organisation wanted to improve sales of a ground-breaking drug that had failed to live up to expectations</td>
</tr>
<tr>
<td>Contributed to ultimately identifying drivers of performance leading to significant impact</td>
<td><strong>Pharma</strong></td>
</tr>
<tr>
<td>6 years of latent data from 14 systems</td>
<td><strong>Data source and 140 variables</strong></td>
</tr>
<tr>
<td>92+ TB of data</td>
<td>Analysed, ingested, and linked different data sources in 8 weeks which resulted in the creation of 140 variables, providing more comprehensive input to the models</td>
</tr>
<tr>
<td>13% increase in project profitability</td>
<td><strong>Reusable pipeline</strong></td>
</tr>
<tr>
<td>Aggregated data covering 12,000 projects to evaluate the variance of project profitability and cost across the three business divisions</td>
<td>That was handed over to the client, and helped empower the client to use internal and external data sets for future analysis</td>
</tr>
<tr>
<td>Analyzed to quantify the value of projects by fee size</td>
<td><strong>Protocol and data model</strong></td>
</tr>
<tr>
<td>Contributed to ultimately identifying drivers of performance leading to significant impact</td>
<td>Developed for exploiting external data sets and automating variable creation, which can be reused across other pharmaceutical use cases</td>
</tr>
</tbody>
</table>
If you want to hear more about Data-Engineering
Join the Meetup DE4DS
Data Engineering Meetup