Data science and AI applications in eCommerce

Dr. Shen Liu, Principal at Logickube

2 August 2023
Agenda

- Introduction
- Case study 1 - Decision Engine
- Case study 2 - Vertex AI
- Case study 3 - Cloud Functions & Bigquery
- Q&A
Introduction
An overview of Logickube

We partner with leading cloud providers to provide AI and Data services to leading retailers in APAC

**Strategic advisory on AI**
Build out a strategic roadmap and Proof-Of-Concepts models in collaboration with our technology and data experts to accelerate your AI journey.

**Cloud data engineering**
Maximise the advantages of new technologies in the cloud by modernizing your data and ML workloads in a scalable and secure manner.

**AI engineering**
Build best in class data and model pipelines to produce your own state of the art AI capabilities across retail, digital media, financial services and health.

**Personalisation & Attribution**
Tailor unique interactions for each customer, across marketing channels and store experiences. Quantify personalisation benefits with robust experiment design and attribution.

---

Google Cloud Partner
Databricks Partner Connect
AWS
Azure
Maths/Statistics is the cornerstone

Designing the right model to solve the right problem
- A systematic process to identify and define problems
- Always design the most suitable solution for each specific problem

Interpreting models and creating actionable insights
- Make black boxes transparent
- Translate maths/stats into business language

Experimental design / getting the right data
- Validate maths/stats tools in a real-world context
- Disentangle factors that jointly contribute to business success

Innovation and R&D
- Develop novel, mathematically sound methods adapting to ever-changing business needs
Cloud computing is our go-to skillset

Leveraging cloud computing could significantly boost the performance of data and AI products

- **Data availability and reliability** - Data are replicated and stored in different locations, easy to backup and restore data in case of any failure
- **Big data capability and high efficiency** - e.g. we use Databricks to optimise performance, enabling real-time data processing, message producing and delivery
- **Cost effective**
  - Reduced cost of maintaining hardware and software
  - ‘Pay-as-you-go’: cost is only generated for what/when is used
  - Various tiers of computing power and storage classes

Cloud based solutions provide a good level of **data security**

- Advanced security features ensure data is securely stored and handled
- Data encryption in transit and at rest
- Certain protocols may be enforced to strengthen security

Our team has relevant **experience** and **qualifications**

- Certified data scientists, data engineers, machine learning engineers and solution architects across mainstream cloud platforms
- Highly experienced in building scalable cloud solutions to create end-to-end data and AI solutions
Case study 1

Decision Engine
Decision Engine offering

ML decisioning of offers
- 8+ basket offer constructs
- 10+ category offers
- Hundreds of discount depths

Superior performance
- +25% inc. sales
- x2 redeem rates

Omni-channel experience
- Reach non-email marketable audience
- 20X increase in return on ad spend in paid channels

Multi-objective decisioning
- Optimisation engine that can balance multiple objectives
- e.g.: costs, audience size, returns

Improves operating efficiency
- Always-on and automated pipeline free up execution resources
- Marketers can focus on offer and creative designs whilst engine curates the best action

Framework for test & learn and enabling measurement
- Rapidly test out new offers through random experiments
- Set aside control groups for measurement purpose
Decision Engine key features

Responsive and personalised recipes across discovery pages
- Home page
- Landing page
- Listing page

Learn all metadata on a recipe, for 5K+ recipes

Real time API
- <200ms latency
- Platform agnostic

Benefits
- Increased engagement and discovery
- Combines offline and online events
- Responsive content
- Automate recipe tagging

4 personalisation API

Popularity
- Most recent
- What's trending
- We think you'll like
Case study 2

Vertex AI
Vertex AI

- Unified development and deployment platform for data science and machine learning
- Increase productivity of data scientists and ML engineers
What Data Science & Machine Learning Engineering teams want

- **Accelerate time to market**
  - Unified data and AI platform for all users to accelerate time to value
  - Vertex AI

- **Enhance stability and reliability of ML systems**
  - End-to-end MLOps to efficiently and responsibly manage and govern AI
  - Vertex AI

- **Improve AI team productivity**
  - Open and scalable AI infrastructure to flexibly and successfully deploy AI
  - Vertex AI
Vertex AI is a platform for all users throughout the ML lifecycle

**Data analyst**
Query and analyse

**Data engineer**
Get clean, useful data

**Data scientist**
Models that work

**ML engineer**
Models in production

---

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data analyst</th>
<th>Data engineer</th>
<th>Data scientist</th>
<th>ML engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endless EDW</td>
<td>Data analyst</td>
<td>Data engineer</td>
<td>Data scientist</td>
<td>ML engineer</td>
</tr>
<tr>
<td>Self-driving infra</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad choice of tools/language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data quality /lineage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real-time capabilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portable notebooks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model eval and selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point-and-click dev</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalable model hosting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning in SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endless EDW</td>
<td>Data analyst</td>
<td>Data engineer</td>
<td>Data scientist</td>
<td>ML engineer</td>
</tr>
<tr>
<td>Self-driving infra</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad choice of tools/language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data quality /lineage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real-time capabilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portable notebooks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model eval and selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point-and-click dev</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalable model hosting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning in SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Vertex AI for large enterprises
Vertex AI for large enterprises

- Vertex AI enables seamless connections with data sources such as BigQuery and Cloud Storage
- Less infrastructure configuration
- End-to-end ML workflows in one place
Vertex AI Workbench: One-stop surface for data science

**Fully managed compute with admin control**
A Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with easily enforceable policies and user management.

**Fast workflow for data tasks**
Seamless visual and code-based integrations with data & analytics services.

**At-your-fingertips integration**
Load and share notebooks alongside your AI and data tasks. Run tasks without extra code.
Benefits

**Easy data exploration and analysis** with Easy access to data in BigQuery and Cloud Storage within a Jupyter notebook

```
Python
# Bigquery regions_by_country
SELECT
    country_code,
    country_name,
    COUNT(DISTINCT region_code) AS num_regions
FROM
    'bigquery-public-data.google_trends.international_top_terms'
WHERE
    refresh_date = DATE_SUB(CURRENT_DATE, INTERVAL 1 DAY)
GROUP BY
    country_code, country_name
ORDER BY
    num_regions DESC;
```

Query complete after 0.19s: 100% 4/4 [00:00=00:00, 16.35rows/s]

```
Python
# regions_by_country.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41 entries, 0 to 40
Data columns (total 3 columns):
   #   Column       Non-Null Count   Dtype
---  ------       --------------   -----  
   0   country_code  41 non-null    object
   1   country_name  41 non-null    object
   2   num_regions   41 non-null    int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

**Fast prototyping and model development** by creating a new notebook under 1 minute and connecting to other GC services within it

```
Python
for i in range(4):
    subprocess.run(['sys.executable', 'AKL_DIR/eval_fit.py', i], shell=True)
```

```
User-Managed Notebook

User-managed notebooks are high customisable VM instances and suitable for data exploration, analysis and model development.

A preinstalled suite of ML/DL packages

- **Python 3**
  - Includes scikit-learn, pandas and more
- **Python 3 (CUDA Toolkit 11.0)**
  - Optimized for NVIDIA GPUs
- **TensorFlow Enterprise**
  - Includes Keras, scikit-learn, pandas, NLTK and more
- **PyTorch 1.13**
  - Includes scikit-learn, pandas, NLTK and more
- **R 4.2**
  - Includes basic R packages, scikit-learn, pandas, NLTK and more
- **Kaggle Python [BETA]**
  - Python image for Kaggle Notebooks, supporting hundreds of machine learning libraries popular on Kaggle
- **JAX 0.3.14 [EXPERIMENTAL]**
  - Data science framework for machine learning research and optimizing runtime for scientific computations
- **Smart Analytics Frameworks**
  - BigQuery, Apache Beam, Apache Spark, Apache Hive and more

Similar setup process to GCE

- **Details**
  - **Environment**
    - **Operating system**
      - Debian 10
  - **Machine type**
  - **Disks**
  - **Networking**
  - **IAM and security**
  - **System health**

**Environment**

All environments use JupyterLab 3 by default and have the latest NVIDIA GPU and Intel libraries and drivers installed. You can specify a previous version instead. Learn more

- **Version**
  - Use the latest version
  - Use a previous version

**Metadata**

Some metadata keys including enable-oslog, framework, notebooks-api, nvidia-driver-gpu-paths, proxy-uri, restriction, shutoff-del-script, title, version are reserved for system use only. If you use these variable names below, they will be overwritten by system values.
## Managed Notebook vs User-Managed Notebook

<table>
<thead>
<tr>
<th>Feature</th>
<th>Managed Notebook</th>
<th>User-Managed Notebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexibility</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Custom environment</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Switch machine type</td>
<td>Within JupyterLab</td>
<td>Shutdown, switch, and restart</td>
</tr>
<tr>
<td>GCS navigation</td>
<td>Within JupyterLab</td>
<td>In GCS</td>
</tr>
<tr>
<td>BigQuery navigation</td>
<td>Within JupyterLab</td>
<td>In BigQuery</td>
</tr>
<tr>
<td>Scheduled runs</td>
<td>Supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>Management fees</td>
<td>$0.05 per vCPU per hour</td>
<td>$0.005 per vCPU per hour</td>
</tr>
<tr>
<td>Idle shutdown</td>
<td>Supported</td>
<td>Not supported</td>
</tr>
</tbody>
</table>
Vertex AI **Pipelines** orchestrate ML workflows serverlessly, and automate and monitor repeatable workflows such as model training and production.

**Benefits:**
- Serverless service
- Lower costs
- Workflow automation
- Composable and reusable pipelines
- Python function-based components
When to use Pipelines

1. **Train/productionise models with well-defined and reusable workflows**
   
   When ML workflows are finalised and will be reused for multiple times, consider packaging the dependencies into a Docker image and migrate the workflows from notebook to Pipelines to save time and improve reliability.

2. **Automate model training/production**

   Manual weekly/monthly model scoring or refitting could be tedious, and schedule pipeline execution or trigger pipeline runs with Pub/Sub could be a game changer.

3. **Scalable model production**

   Built on top of Kubernetes, Vertex AI Pipelines are serverless and scalable. Users are able to specify different level of resources for different steps and design parallel processing to boost speed.
Automate Model Training/Production

Model training/production can be automated by scheduling or triggering pipeline runs.

To schedule pipeline runs, **Cloud Scheduler** and **Cloud Functions** are also needed other than Vertex AI:

- Configure Cloud Scheduler to send a JSON string to Cloud Functions on your pre-defined schedule.
- Cloud Functions that you build will parse the JSON string and submit pipeline runs using ingested parameters.
- Pipeline runs

![Diagram showing the flow of Cloud Scheduler to Cloud Functions to Vertex AI Pipelines]
Case study 3

Cloud Function & Bigquery
A Multi-agent Orchestration Problem

- Multiple teams in the company collaborate for a common business (orchestration);
- They have different preferences in data transfer methods (Gmail, RDMS, Google Drive);
- They use different technical tools (GCP, Azure, AWS);
- They own different domain knowledge (BI, DA, DS);
Solution to the Orchestration Problem

Serverless Architectures
● Faster solutions to market at lower cost
● Decreased management overhead than traditional approaches

Cloud Function v.s. Cloud Run

● Cloud Function
  ○ Transforming data and loading it into BigQuery
  ○ Creating data summary once a BigQuery table gets updated
  ○ Use ML APIs to analyze data added to a database or storage bucket

● Cloud Run
  ○ Any web-based workload
  ○ REST APIs for mobile apps or games
  ○ Internal custom backoffice apps

Google Cloud Function
● Function-as-a-service (FaaS) in Google Cloud;
● Serverless architectures with pay-as-you-go convenience;
● Connection or extension to services with complex applications;
● Remedy to reconcile orchestration problems;
Main Features of Google Cloud Function (GEN 2)

**Increased compute with granular controls**
- Instance concurrency (up to 1000 requests/instance)
- Fast rollbacks (version control)
- 6x longer request processing (max. 60 minutes)
- 4x larger instances (max 16GB RAM + 4 vCPUs)
- Pre-warmed instances (fast configuration)
- Support multiple programming languages
- Extensibility and portability (to Cloud run)

**Empowering Business Intelligence**
- Inclusive to contributors from different backgrounds;
- Enable non-SQL functionalities;
- Combine complex operations in one go;
- Seamless data/messages digestion + broadcast;

---

**Runtime**

<table>
<thead>
<tr>
<th>Runtime</th>
<th>Operating System</th>
<th>Runtime ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python 3.11</td>
<td>Ubuntu 22.04</td>
<td>python311</td>
</tr>
<tr>
<td>Python 3.10</td>
<td>Ubuntu 22.04</td>
<td>python310</td>
</tr>
<tr>
<td>Python 3.9</td>
<td>Ubuntu 18.04</td>
<td>python39</td>
</tr>
<tr>
<td>Python 3.8</td>
<td>Ubuntu 18.04</td>
<td>python38</td>
</tr>
<tr>
<td>Python 3.7</td>
<td>Ubuntu 18.04</td>
<td>python37</td>
</tr>
</tbody>
</table>

Node.js, Go, Java, Ruby, PHP, .NET Core
Main Features of Google Cloud Function (GEN 2)

Lots more event sources with the Eventarc

- 125+ Event sources (BigQuery, GCS, API Keys)
- Standards-based Event schema for consistent developer experience
- Customer-Managed Encryption Keys (CMEK) support
Orchestration Connected by Cloud Functions

Simple ETL Workflow

Transfer Video Objects to GCS

Real-time Text Messages Recognition and Logging
Comparison to Conventional Orchestration Pipeline

- BQ Table 1
- BQ Table 2
- BQ Table 3
BigQuery ML Overview

```sql
#standardSQL
CREATE MODEL `bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, '') AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, '') AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20160801' AND '20170630'
```

BigQuery ML is a part of enterprise BigQuery that allows you to create and execute ML models using Google SQL queries.
Why Use BQML?

- **Easy adaptation**
  - Develop ML models using the language you are comfortable with
  - No need to learn Python or Java and ML frameworks such as TensorFlow or PyTorch

- **Increased development speed**
  - No need to move data in/out of BQ throughout the entire ML lifecycle.
  - Bring ML to data, not the other way around.
  - No need to wait for limited resources of data science team

- **No more time wasted on setup**
  - BigQuery is serverless so no need to provision VMs for model training
  - Ready to develop - no extra setup required such as installing frameworks and other dependencies
## BQML - Supported Models

### Internally trained

<table>
<thead>
<tr>
<th>Regression</th>
<th>Classification</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>Logistic regression</td>
<td>K-means clustering</td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td>Matrix factorisation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time series forecasting</td>
</tr>
</tbody>
</table>

### Externally trained (Vertex AI)

<table>
<thead>
<tr>
<th>Regression</th>
<th>Classification</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>Wide &amp; Deep Networks</td>
<td>Autoencoder</td>
</tr>
<tr>
<td>Wide &amp; Deep Networks</td>
<td>Boosted Tree</td>
<td>AutoML Tables</td>
</tr>
<tr>
<td>Boosted Tree</td>
<td>Random forest</td>
<td>AutoML Tables</td>
</tr>
<tr>
<td>Random forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AutoML Tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autoencoder</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BQML in Google ML Landscape

Out of box

Pre-trained APIs & solutions
Cloud Vision API
Speech-to-Text API
...

Custom AI with BQML and AutoML
No-code/low-code approach

End-to-end AI with core tools
Vertex AI and TensorFlow give data scientists strong control to build and deploy models

DIY
You can import the following models trained outside BQML and use them to perform prediction within BQ:
- Open Neural Network Exchange (ONNX) format
- TensorFlow Saved Model format
- TensorFlow Lite format
- XGBoost Booster format

You can export most models trained in BQML in the following formats and use them in other environment:
- TensorFlow Saved Model format
- XGBoost Booster format
Integrate BQML in Vertex AI

Advanced BQML models are usually trained in Vertex AI, which is Google's unified ML platform.

Integrating BQML in Vertex AI gives you online model serving capabilities and allows you to manage BQML models just like any other ML models via Model Registry.
Typical machine learning workflows

**Training/Creation Phase**
- **Labelled Data**
- **ML Algorithm**
- **Trained Model**

**Inference/Use Phase**
- **New Data**
- **Served Model**
- **Predictions**

---

**Steps**
- Objective identification
- Data preparation
- Model building
- Model training
- Model evaluation
- Model deployment & maintenance
Create a BQML model using CREATE MODEL

```
#standardSQL
CREATE OR REPLACE MODEL `bqml_demo.penguins_model`
OPTIONS
(model_type='linear_reg',
  input_label_cols=['body_mass_g']) AS
SELECT *
FROM `bigquery-public-data.ml_datasets.penguins`
WHERE body_mass_g IS NOT NULL
```
Evaluate a BQML model

Evaluation is often automatically done during model creation in BQML, to early stop the model training process to avoid overfitting.

The validation set is used in this process, so it is also known as validation.
Use a BQML model using ML.PREDICT

Use your trained model to make predictions on new data, e.g., in model production.

```
SELECT *
FROM ML.PREDICT('bi-workshop-2023-92764.bqml_demo.website_conversions')
    SELECT
        IFNULL(device.operatingSystem, '') AS os,
        device.isMobile AS is_mobile,
        IFNULL(geoNetwork.country, '') AS country,
        IFNULL(totals.pageviews, 0) AS pageviews,
        IFNULL(totals.timeOnSite, 0) AS time_on_site
    FROM `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'
```

Query results:

<table>
<thead>
<tr>
<th>Row</th>
<th>predicted_label</th>
<th>predict_label</th>
<th>predict_prob</th>
<th>os</th>
<th>is_mobile</th>
<th>country</th>
<th>pageviews</th>
<th>time_on_site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.02068778479696</td>
<td>BlackBerry</td>
<td>true</td>
<td>Indonesia</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0.00177014569063</td>
<td>BlackBerry</td>
<td>true</td>
<td>Indonesia</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0.015617142330</td>
<td>Samsung</td>
<td>true</td>
<td>India</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0.01443974952</td>
<td>Windows Phone</td>
<td>true</td>
<td>India</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0.014605231870</td>
<td>Windows Phone</td>
<td>true</td>
<td>Romania</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Contact

Shen Liu
s.liu@logickube.com
www.logickube.com
https://www.linkedin.com/company/logickube