Predicting job openings based off the Covid-19 infection rate

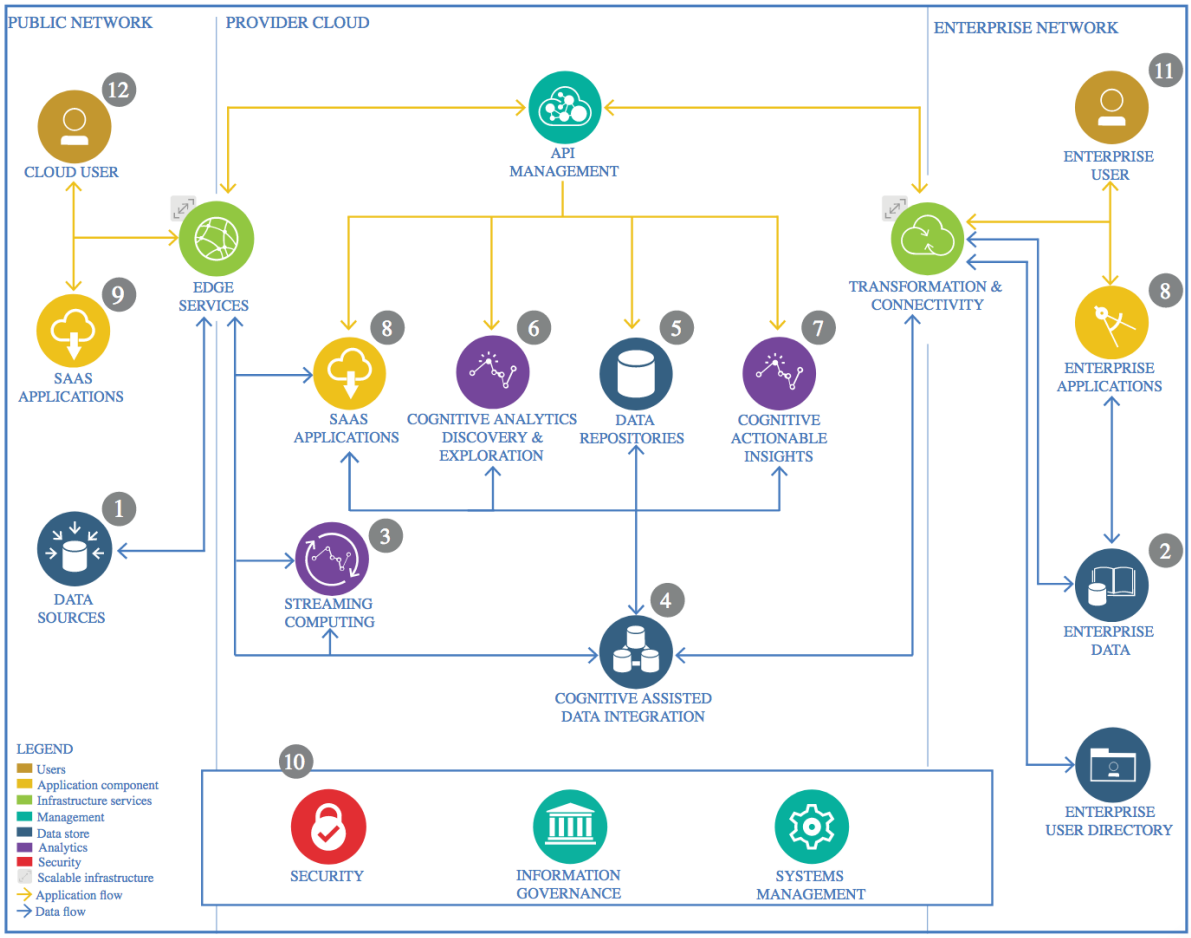
Artificial Decisions Document

IBM Advanced Capstone Project

15/01/2020

Arron Hovingham

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

In our IBM advanced data science capstone project, we will be making use of two distinct data sources.

Firstly, we will be making use of the online job adverts estimates dataset that is provided by Adzuma for the Office of National Statistics. This dataset, which we have linked to below, provides information on the number of jobs openings in the UK advertised across different sectors and regions.

https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/onlinejobadvertestimates

We will then be using the UK Covid-19 infection rate dataset which is also provided by the Office of National Statistics. This dataset provides information about the Covid-19 infection rate across different parts of the UK and can be accessed through the link below:

https://coronavirus.data.gov.uk/details/cases

1.1.1 Justification

As an originally devised project, these two datasets are all we will need in order to generate a sophisticated end data product. Being open sourced datasets as well, both these datasets are free for us to access, use, and modify for the entire duration of our project.

1.2 Enterprise Data

In this project, we do not plan on using any enterprise data.

1.2.1 Justification

As all the data we are going to be using in our project is open sourced and publicly available, we will not be incorporating any enterprise data within our project.

1.3 Streaming analytics

In this project, we do not plan on using any streaming analytics software.

1.3.1 Justification

As our project is not making use of any live event data, incorporating streaming analytics software within our project would not provide any benefit to us.

1.4 Data Integration

To integrate our two data sources together, we will be using the Pandas package in Python.

1.4.1 Justification

Pandas is a very easy and simple to use open source Python package that will allow us to easily integrate our two datasets together; alongside this, an additional benefit of us using Pandas is that it also allows us to easily manipulate our datasets when selecting features, handling any potential missing values, or addressing any data quality issues within our datasets.

1.5 Data Repository

For our project, we will be using GitHub as our primary code storage repository. We will also be making use of the IBM cloud as a backup cloud storage option.

1.5.1 Justification

GitHub is a safe and secure cloud-based coding storage repository and the perfect environment for us to store our datasets and code base. We will also be storing our codebase on the IBM cloud, so that if needs be, we can incorporate the IBM AI platform into our project.

1.6 Discovery and Exploration

To perform our data discovery and exploration processes, we will be using the pandas, matplotlib, and seaborn Python libraries within Jupyter Notebooks.

1.6.1 Justification

Jupyter Notebooks is a fantastic coding development environment that is incredibly simple and straightforward to use. A great advantage of using Jupyter is that it can generate inline visualizations while processing our code, making it an ideal environment for showcasing the findings of our data discovery and exploration processes. On top of this, the three Python libraries will provide us with all the tools we need to perform our discovery and exploration processes.

1.7 Actionable Insights

When we generate our predictions using machine learning, we will be making use of Scikit-Learn package. In addition to this, we will also be integrating TensorFlow libraries into our Jupyter Notebooks so that we can create and develop deep learning models.

Specifically, we will follow the procedure outlined below:

Model Definition

We will be testing four algorithms in this project: Linear Regression, Random Forest, K-Means, and Deep Neural Networks.

Model Training

Our models will be training within Jupyter Notebook using our local computer.

Model Evaluation

There are three methods we will be using to evaluate the effectiveness of our models: Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and R-Squared (R2)

Model Deployment:

Lastly, we will be deploying our models by generating a PDF report using Jupyter Notebooks.

1.7.1 Justification

In this project, we have chosen to utilize a variety of machine and deep learning algorithms. This is because we want to test a range of different approaches to find out which one will perform best in predicting the number of job openings based off the UK Covid-19 infection rate. Alongside this, we have also used a range of different regression evaluation metrics to gain a comprehensive overview of how well each of our models is performing. Finally, we have decided to present our findings in a PDF report format because it is by far the easiest and simplest method of model deployment.

1.8 Applications / Data Products

The aim of our project is to have generated a sophisticated end data product that predicts the number of job openings in the UK based off the UK Covid-19 infection rate.

1.8.1 Justification

We have set out to investigate whether there is a correlation between the Covid-19 infection rate and the number of job opportunities within the United Kingdom. We argue that this is a good question to investigate as it is specific, measurable, achievable, realistic, and can be completed in a reasonable time frame. If any correlation is identified, we can then proceed to develop an advanced learning model which predicts the number of job openings based off the current infection rate.

1.9 Security, Information Governance and Systems Management

We will be storing our datasets in csv files locally on our personal computer.

1.9.1 Justification

Because our datasets are open sourced, comply with GDPR principals, and contain no personally identifiable data: storing these datasets locally on our personal computer will be adequate for the duration of the project.