OCI | WE ARE SOFTWARE ENGINEERS.

Data Engineering

The missing link in operationalizing Analytics & Data Science Initiatives

Data Engineering

The missing link in operationalizing Analytics & Data Science Initiatives

OVERVIEW

Many of our clients are investing heavily in their Analytics and Data Science efforts and teams. These teams are skilled in analyzing and uncovering valuable information in data; but we've found that often, these teams do not have skills in software engineering to deploy and operationalize a secure, maintainable, and scalable application. OCI has implemented a variety of solutions to help your Analytics and Data Science professionals transfer their insights to internal and external constituents in a secure, maintainable, and scalable application.

REAL LIFE EXAMPLE

A Global Life Sciences client was generating prescriptive analytics as part of a consumer facing product leveraging advances in deep machine learning techniques. The production schedule of the application included a trial period to test alternative prescription methods. As the application approached production, the teams worked to build backend components to support a repeatable analytics discovery and deployment platform.

Unfortunately, the client began to see that they needed assistance in developing and integrating the components of their platform.

The client knew they needed to automate the analysis, but they were having difficulty:

1. Getting their data efficiently from their cloud and on-premises data stores
2. Scaling the analysis

Our expert Architects and Engineers were called in to help define and develop the components of a full lifecycle analytics platform that supported the client with:

- A preprocessing pipeline that assembles and QCs the data for analysis,
- Training their deep learning models at scale,
- Deployment of machine learning models for production, and
- Deployment of secure, scalable APIs

Our team developed tools to assist the client in the orchestration of the preprocessing pipeline, including transfers to and from their cloud data store, training of neural networks, and model execution for future scalability. With this design, the client could focus on the data science while the our engineers focused on data retrieval, storage, training, and processing efficiency.
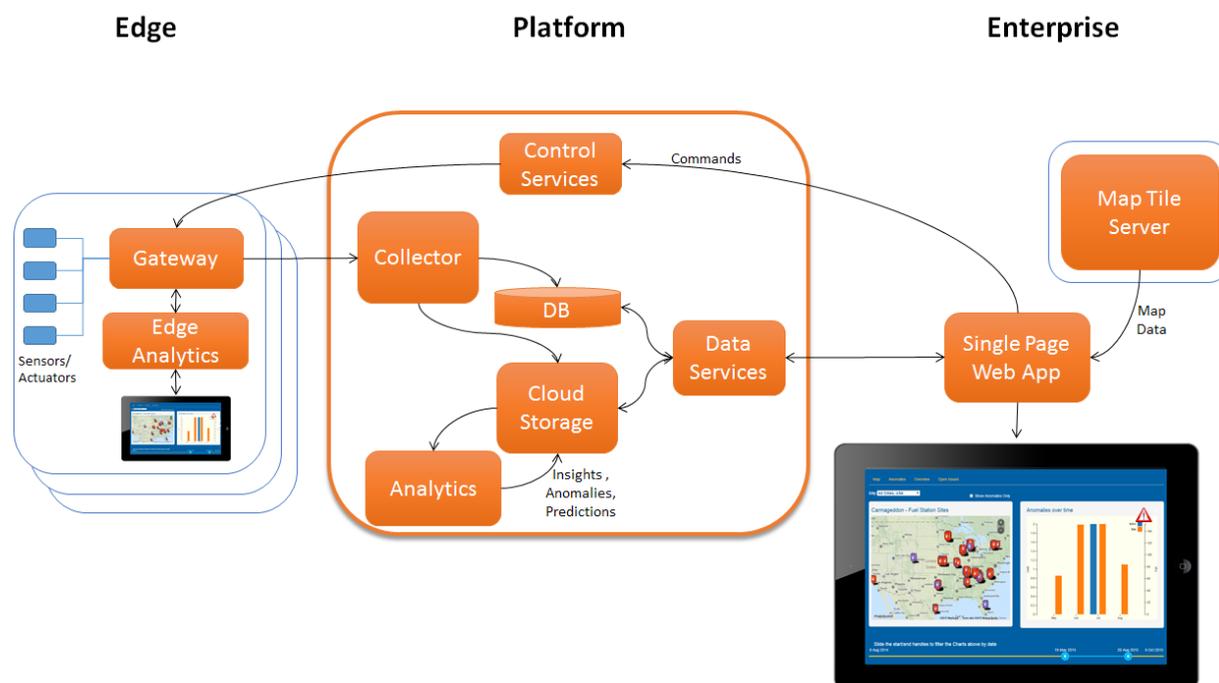
OCI Sample Production Reference Architecture

EXPERTISE IN DIFFERENT DOMAINS - A MULTIPLIER EFFECT

It's hard to argue that Data Scientists are not software engineers and most data science teams could benefit from a software engineering influence. There can be multiple roles necessary in the application of analytics and data science, but we see three primary roles:

1. **Business Analyst** - coordinates between data scientists and executives to help drive decisions based on data. Often works with sales and marketing, utilizing statistics to arrive at insights about the analysis.
2. **Data Scientist** - conducts research and works on identifying opportunities from data utilizing data mining, statistics, pattern recognition, clustering, and other techniques to convert data into knowledge.
3. **Data Engineer** - ensures efficient capture, processing, and storage of data as it flows through the analytics pipeline. Typically has a strong software engineering background and can work well with data scientists to optimize their algorithms and implement solutions that are accessible to the business analyst.

We believe there are three main areas where a Data Engineer can impact data science success:

1. **Programmer efficiency:** Applying software engineering principles can help data scientists be more productive by developing correct, performant, and maintainable software.
2. **Repeatability:** Applying automation techniques allows for repeatability in the scientific process.
3. **Scalability:** A professional software engineer can guide a design into a scalable solution. This allows a data scientist to ask bigger questions and get more accurate and relevant answers.

SOLUTION

We have assembled a pool of Data Engineers that can be used to augment the various data science and business analysis teams within your organization. These Data Engineers bring expertise to bear on the software engineering challenges faced by data science teams to support all phases of a project.

1. **Initial planning:** The Data Engineer is there to fill in the technical details that lie between an idea and its realization in a production system. Specifically, the Data Engineer can provide insight into the available technologies and techniques, potential problems, etc.
2. **Discovery:** Many projects will begin with a proof of concept (POC). The Data Engineer is there to help with data access, preprocessing, and analysis. The engagement may range from high-level discussions to writing code.
3. **Planning:** A successful POC calls for a concrete plan for scaling the POC. This should include the business (the customer), the data science team, and the engineering team that will deploy and manage the product. The role of the Data Engineer is to bring these three parties together from a technical perspective to ensure feasibility and cost effectiveness.
4. **Development:** The Data Engineer will assist with the development of the product through planning, coding, testing, automation, etc. The Data Engineer can provide significant value in preparing the product for production.
5. **Production handoff:** The Data Engineer is there to facilitate the deployment of the product in non-prod and prod environment.

HOW DO I ENGAGE A DATA ENGINEER?

1. **Project management:** Data science teams will have a (project) manager that has the ultimate say in how the project proceeds. The role of the Data Engineer is to provide technical insight to the manager. For example, suppose that the team must acquire data from a system of which they have no experience. The Data Engineer will help form reasonable time estimates and expectations with regard to accessing the new system.
2. **Development:** The Data Engineer can assist with the development of code.

3. **Automation:** Automation is critical for repeatability in data science and also helps scalability. The Data Engineer is there to provide both thought leadership and practical expertise.
4. **Teaching/Coaching:** Data Engineers take various technologies and techniques for granted, e.g., source code management (git), continuous integration (Jenkins), and project management (JIRA), etc., code review, Agile practices, etc. However, many data scientists are not familiar or comfortable with these technologies and techniques. Thus, the Data Engineer has a significant opportunity to educate data scientists and promote personal development. This could come in the form of coaching, code reviews, or formal classes.

Let OCI provide you access to highly qualified Data Engineers who have experience in project management, development, scalability, automation, and who know how to leverage and build upon the skills of your existing or emerging data science team. Contact us today to speak with a Data Engineer about your project.

info@objectcomputing.com or +1 (314) 579-0066