Big Data

Big Data Infrastructure, Analytics, Exploration and Visualisation

We have extensive experience in Big Data solutions, and particularly with Google Cloud Platform (GCP) hybrid Infrastructure as a Service (IaaS) / Platform as a Service (PaaS), such as those displayed in the Canaccord Quest Case Study.

Big Data is inseparable from Infrastucture, and is useful only in the context of Machine Learning, both of which we are experts at.

Collecting large amounts of data requires appropriate scalable infrastructure. And collecting data is only worthwhile if it is analyzed, processed, and visualized.

Technologies and Tools

Big Data Management

- Apache Spark: Fast and large scale data processing engine for real-time analytics, batch processing, Interactive analytics, and graph processing. Runs on resource managers like Mesos, YARN.
- Hadoop: Distributed storage and processing platform for large data sets. Data crunching is done over YARN cluster management technology and batch processing framework MapReduce.
- Apache Flink: Distributed Stream and batch processing engine. Provides data distribution, communication, and

fault-tolerance for distributed computations over data streams.
- Storm: Distributed real time computation system which reliably processes unbounded streams of data.
- Data Ingestion and retrieval: Continuous ingestion and querying of data with Hadoop ecosystem technologies like Flume, Sqoop, Kafka, Hive, Drill, Elasticsearch
- NoSQL Databases: Storage and retrieval of unstructured and semi-structured data with NoSQL databases like MongoDB, Cassandra, HBase.

Hadoop Alternatives

For a brief review of Open Source Data Lakes alternatives to Hadoop (and Big Query), please read the article "Hadoop Alternatives".

The tools reviewed are:

- Pachyderm
- Apache Spark
- Google BigQuery
- Presto
- Hydra

Data Analytics

- Machine Learning: Scalable and extra ordinary fast machine learning computations with Apache Spark ML, MLlib, Apache FlinkML, Apache Mahout.
- Graph Processing: Iterative and parallel graph computations with Spark GraphX, Flink Gelly.
- Structured big data analytics: SQL like interface and processing of large scale data with Spark SQL, Flink Table APIs.
- Advanced Analytics: Recommender Systems and Predictive Analytics using machine learning algorithms.
- Statistical Computing and Analysis: Use of R, Python,

Scala programming languages for statistical analysis, predictive modelling and validations.

Data Exploration and Visualization

- Tableau: Popular visualization tool which supports wide variety of graphs, charts, maps and other diagrams.
- D3.js: Uses HTML, CSS, and SVG to render amazing charts and diagrams.
- ElasticSearch and Kibana

Infrastructure

On the infrastructure side, we take advantage of the exceptional infrastructure and various PaaS tools available on GCP:

We offer consulting and development Big Data services, partnering with Google Cloud Platform and their array of Big Data hybrid PaaS / IaaS tools such as :

- BigQuery: a fully managed low cost analytics database.
- DataLab: a powerful interactive tool created to explore, analyze and visualize data.
- DataProc: a managed Hadoop MapReduce, Spark, Pig, and Hive service, to easily process big datasets at low cost.

Machine Learning

Once the Data is organized, collected, and made available on large scalable infrastructure, it has to be analyzed, and our expertise in Data Science is at the forefront of modern techniques.

As such, we can offer a comprehensive solution, from

organizing and managing infrastructure, to analyzing your Data to make sense and ultimately benefit from it.

Please read our Data Science case studies, covering the following industries:

- Aerospace (turbulence prediction for airlines)
- Finance (algorithmic trading)
- Insurance (fraud detection)
- Marine vessel route monitoring (anomaly detection for vessels trajectory)
- Search (personalized search via Graph methods, search engine)
- Website Monitoring (courbe de charge pour des sites B2B)