Bigstep DataLab: a Technical Overview

January 2017

bigstep

Contents

Contents 2
Notice 2
Introduction 3
Main Features of Bigstep DataLab 4
  Collaborative Data Exploration 4
  Adaptability 4
  Integration 4
Components of Bigstep DataLab 5
  Bigstep Data Lake 5
  Bigstep Real-Time Container Service 5
  Bigstep Real-Time Spark
 Service 6
  Bigstep SparkSQL Service 6
  Bigstep Kafka Service 6
  Bigstep Zookeeper Service 6
  Jupyter 6
  Zoomdata 6
Bigstep DataLab Architecture 6
Sales Data Analysis Use Case 7
Security and Network Architecture 9
Performance 9
About Bigstep 10

Notice

Please note that the entire information contained in this document is provided by Bigstep and its affiliates on an “as is” basis, exclusive of any warranty whatsoever. Bigstep makes reasonable efforts to always include accurate and updated information in this document, however, no warranties are made regarding its accuracy and completeness and the Client uses this information at their own risk.

Bigstep may periodically update the information in this document without any prior written notice.

Bigstep, its suppliers or any other party involved in creating, producing or delivering this document’s contents are not liable for any indirect, special, incidental or consequential loss or damages arising from the use of the information contained in this document, whether in an action in contract tort (including negligence) or otherwise; any loss of profit; any loss of business or any loss of data arising from the access or use of the information contained herein.

Introduction

Bigstep DataLab is a turn key data-research solution that enables easy access to collaborative analytics and data science. It simplifies access to state-of-the-art software like Apache Spark, Apache Kafka, Jupyter, Zoomdata, and others in order to power data-driven decision making.

The DataLab abstracts away all the complexity usually associated with this type of projects, lowers the cost to a fraction and enables a quick start in new analytics and data science projects without any infrastructure or IT skills required.

It supports the translation of unstructured or semi-structured data into structured data with schema-on-read capabilities. This enables access to a whole new range of data sources such as external SaaS solutions like Salesforce, Google Analytics as well as clickstream records, sensor data and many others.

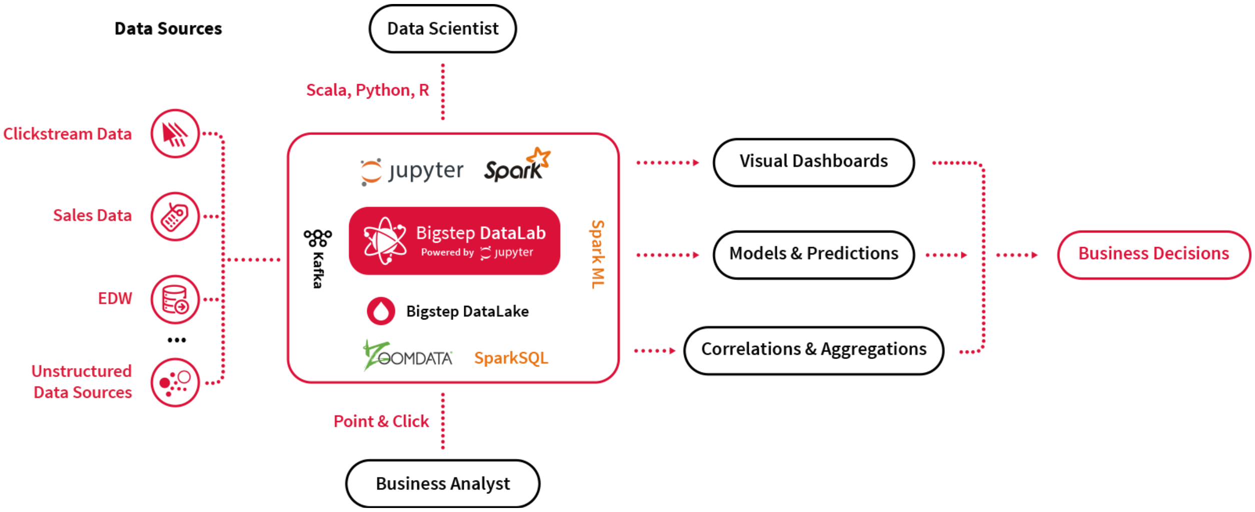

Figure 1: The Bigstep DataLab in Context

Big data investments amounted to 0.6% of corporate revenues and returned a multiple of 2.0 times the level of investment over five years, while profits increased by 9% over the same period, as shown in a 2016 McKinsey study.

However, access to state-of-the-art technologies has proven difficult and high-priced for smaller, less technical companies. The Bigstep DataLab enables access to otherwise expensive technologies and practices by reducing cost, complexity and the level of skills required to operate.

Bigstep DataLab can easily handle large quantities of real-time and historical data, perform complex machine-learning tasks and be quickly stopped or repurposed, with on-demand scalability and pricing. It enables users to experiment with powerful data technologies, leveraging Bigstep's award-winning bare-metal infrastructure.

Main Features of Bigstep DataLab

Collaborative Data Exploration

Cross-functional, data exploration groups, which cumulate data science, domain expertise and business intelligence skills, reporting directly to the CEO have become a regular occurrence in larger enterprises. Conceptually, the DataLab aims to be a single, unified environment, where data from multiple sources is explored collaboratively using the tools that are compatible with each user's technical or non-technical background and skills.

Smaller teams typically rely on cross-functional individuals that have gotten to know a bit of everything. The DataLab aims to save these teams the time and trouble required to set up such a complex software and infrastructure.

Two very important dimensions of the Bigstep DataLab are the Data Science component (e.g. Jupyter, Zeppelin, Spark Streaming) and the Visual Data Exploration component (e.g. Superset, Zoomdata, Qlik, Tableau). They are designed to seamlessly share data via the DataLake while running on the same compute fabric.

Adaptability

By definition any laboratory needs to allow rapid modifications, as goals change over time. The Bigstep DataLab architecture can run multiple Spark clusters side by side on the same underlying hardware.

All components of the DataLab, including the underlying hardware are scalable on demand in order to adapt to varying workloads. Memory is configured automatically when new resources are added to the cluster to avoid out-of-memory errors when dealing with complex data analysis.

The smallest deployment of the DataLab is a single server, for minimal footprint. It can then be expanded on-line to hundreds of servers, to cope with demand.

Integration

The structured data stored in the DataLake is accessible via SparkSQL over JDBC/ODBC. The metadata is stored centrally to all Spark clusters and all tools see the same table list. This allows an external tool such as Qlik or Tableau to access the data.

The unstructured data or the *parquet or avro* files are accessible via a standard Hadoop Filesystem implementation and via a native HDFS and WebHDFS protocols. This can link the DataLake into an existing Hadoop cluster.

Spark has connectors for many data sources including SQL databases such as Oracle, MySQL, PostgreSQL, DB2 and also to NoSQL databases such as MongoDB, Couchbase and others, as well as connectors for SaaS services such as Salesforce.

Data can also be ingested in real-time via the Kafka connector and consumed on the fly, or serialized into the DataLake. Many tools, such as Logstash or Streamsets, can produce data for Kafka. The underlying infrastructure is optimized for real-time processing, having high performance and low latency.

Components of Bigstep DataLab

The components of the DataLab are pre-integrated and hand-picked to serve a specific function for the target audience: Data Scientist, BI, Analysts and Decision Makers. The DataLab is designed to operate without the need to involve IT, but within an environment IT can fully control.

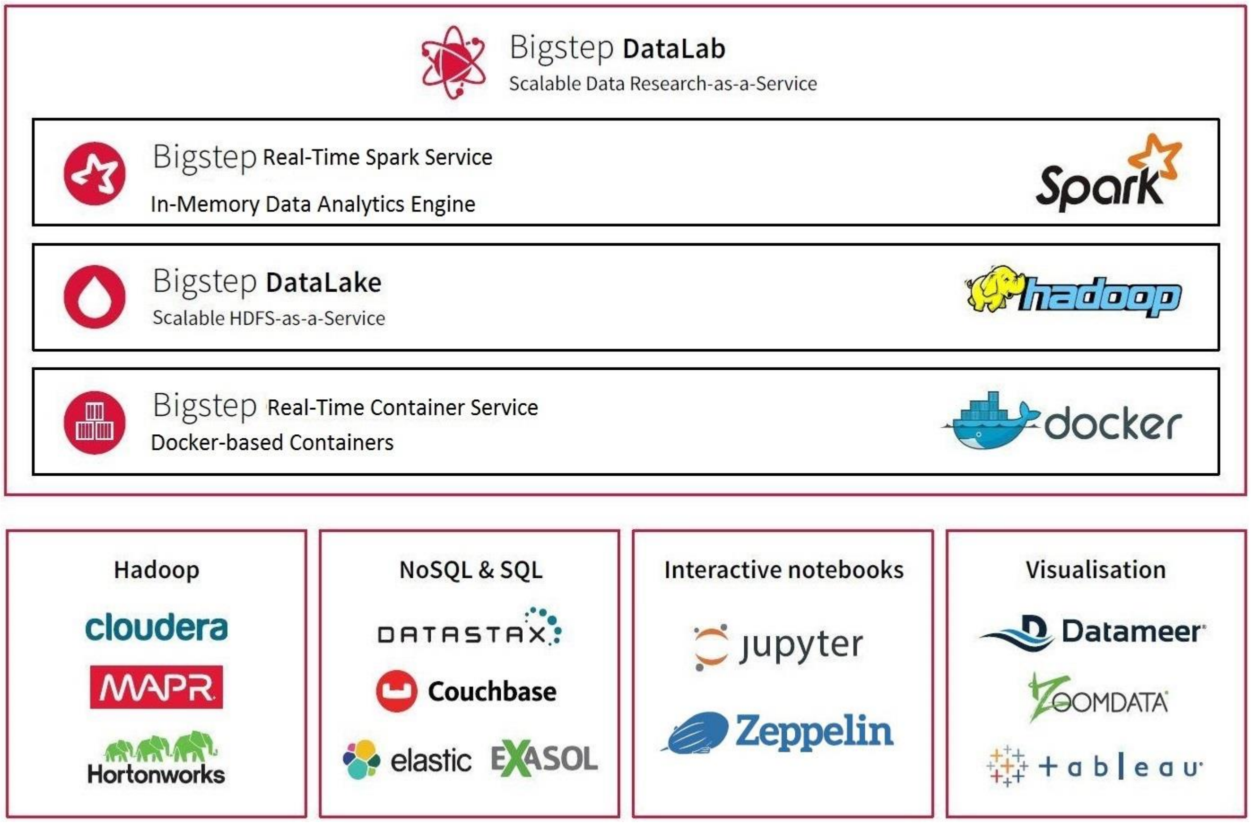

Figure 2: Bigstep DataLab on Bigstep Big Data Platform

Bigstep Data Lake

A scale-out, affordable data repository acting as a “single-source-of-truth” for all the projects within the DataLab. It is fully HDFS compatible, encrypted and resilient.

Bigstep Real-Time Container Service

This is a Docker-based application hosting service that aggregates the resources of multiple bare-metal servers into a single, big server without using virtualization. The applications that run in containers on this large, distributed server, each have dedicated CPU cores, RAM and Disk resources, are fully isolated from one another and run directly “on the metal.”

Bigstep Real-Time Spark Service

A pay-per-use, fully managed, large-scale, in-memory data processing service capable of machine learning and graph processing that leverages Apache Spark. Bigstep uses multiple, independent, smaller clusters for separate jobs, to simplify management in multi-tenant environments.

Bigstep SparkSQL Service

SparkSQL is a service that allows industry standard JDBC and ODBC connectivity for business intelligence tools to data coming from a variety of sources and Spark programs.

Bigstep Kafka Service

Is a fast, scalable, queueing system, designed as an intermediate layer between producers and consumers of data. It can be used to bring data into the DataLab by connecting it via Spark Streaming jobs.

Bigstep Zookeeper Service

Zookeeper is a centralized service for maintaining configuration information, naming, and providing distributed synchronization for distributed applications.

Jupyter

This is the de-facto notebook technology for data scientists, and it is pre-integrated with the Spark engine as well as the standard NumPy, SciPy, matplotlib and other tools for the Python, Scala, and R programming languages.

Zoomdata

Zoomdata is a high-level visualization tool that enables users with less technical background to create interactive dashboards powered by the historical, batch and real-time streaming data from their organization.

Bigstep DataLab Architecture

The solution uses multiple components, each with its unique role, ranging from data replication, encryption, to self-service data editing, etc.

The design goal of the architecture is to allow maximum diversity of applications while also being easy to manage. A secondary goal is to have a scalable and resilient environment that allows workloads to be moved across hosts.

The DataLab relies on Docker containers in the ContainerPlatform to segregate applications from one another, as well as increase their resilience. This also allows the user to run custom applications within the same cluster, without having access to the underlying infrastructure.

The individual services, especially the Spark and Kafka, use a variety of Docker images custom built to function as a cluster for easy scaling and resilience while sharing a centralized metadata service and Kerberos identity for accessing the DataLake.

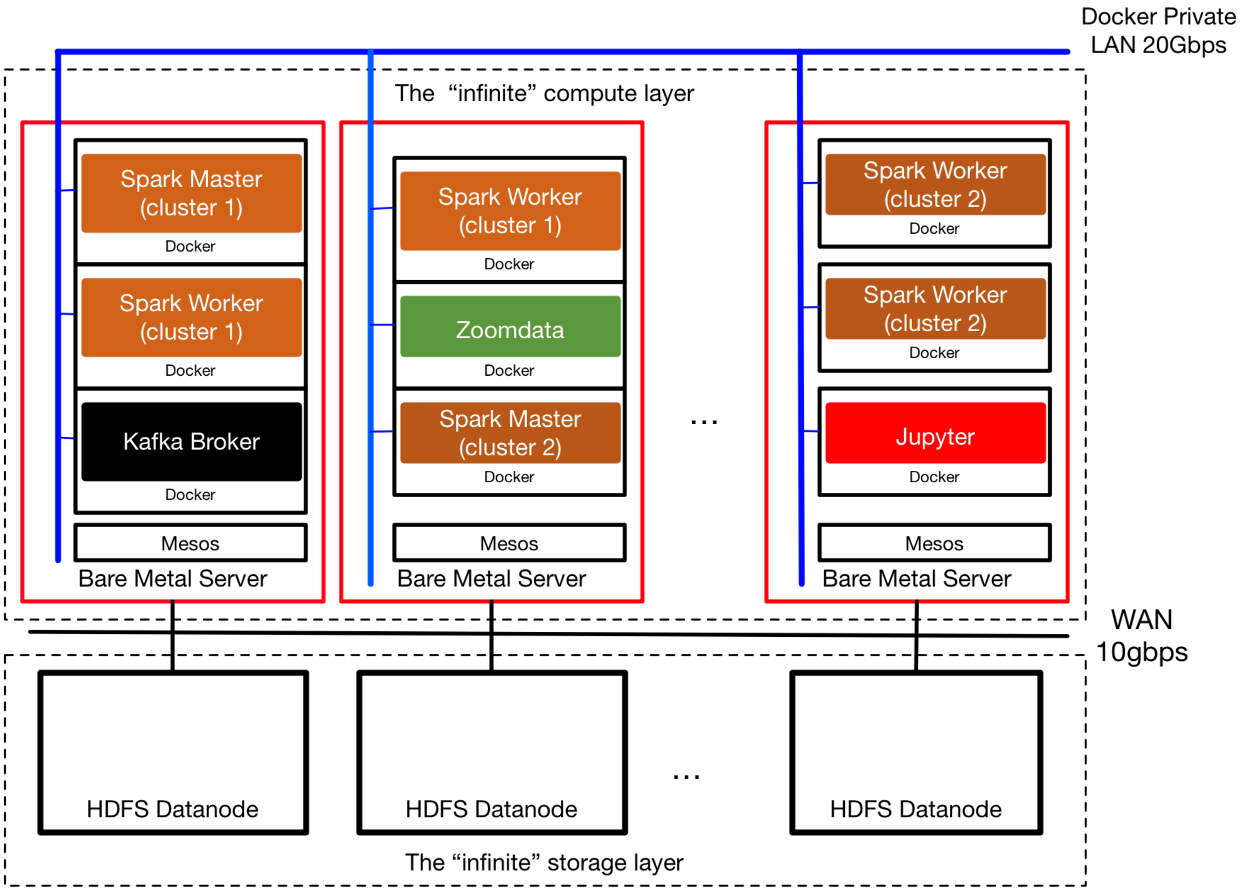

Figure 3: Bigstep DataLab Architecture

Sales Data Analysis Use Case

A sample use case of the Bigstep DataLab would be a sales data analysis pipeline. Depending on the data input, this type of analysis aims to improve customer loyalty, increase ROI, create targeted marketing campaigns, and so on.

Two different flows are required: a ***real-time pipeline*** for data captured from the website traffic, and a ***batch pipeline*** for the extraction and processing of the data stored in an EDW and from Salesforce.

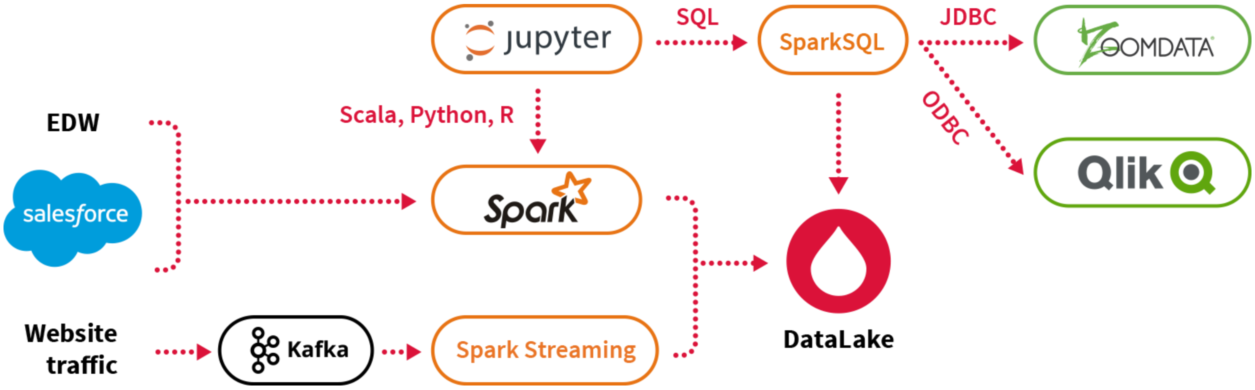

Figure 4: Sample Pipeline Running in the DataLab

The ***batch analysis pipeline*** uses Spark to connect via JDBC to both the EDW and Salesforce (via the DataDirect connector). The code to extract data as well as to process it into aggregated tables is written in Python in a Jupyter notebook by a Data Scientist.
The aggregated data stored in the DataLake and generated by the batch pipeline is accessed using Zoomdata through the SparkSQL ODBC connector by the Business Analyst that creates higher level visual dashboards and reports.

The reports are then used by the Domain Expert to convey the results of the analysis to the decision makers.

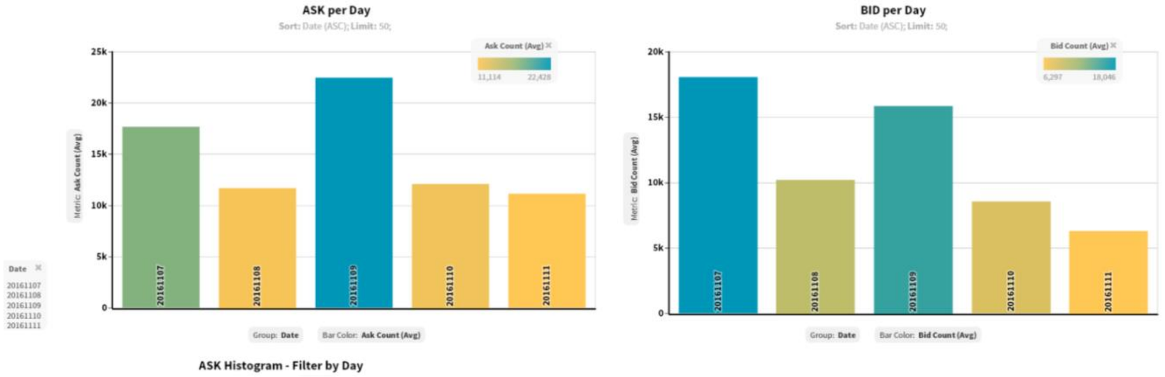

Figure 5.1: Sample Reports Generated with Zoomdata

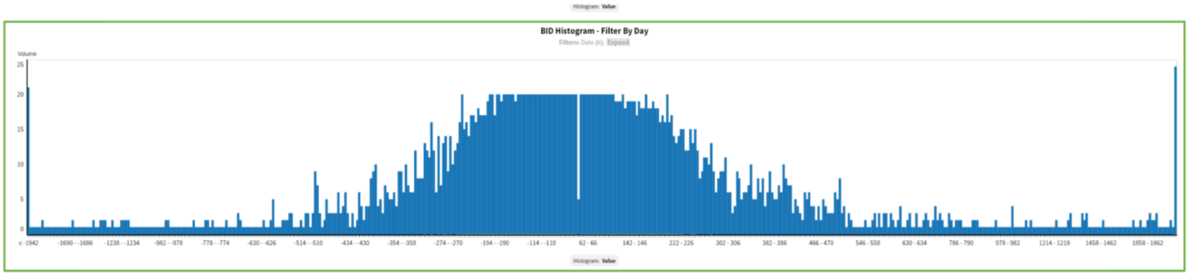

Figure 5.2: Sample Reports Generated with Zoomdata

For the ***real-time pipeline***, data is submitted from the website via Logstash Kafka Producer to the Kafka cluster. The data is then consumed using a Spark Streaming job running on the Spark cluster, cleaned up and serialized in the Bigstep Data Lake as Avro files.

Security and Network Architecture

The DataLab runs within the Bigstep Metal Cloud, on resources abstracted by the Real-Time Container Platform. This abstraction also segregates each individual component from the others in containers that serve a single function. The Docker images are custom built using a minimal number of components, so that attack footprint is reduced.

An automated firewall secures all access to the applications from the Bigstep DataLab. The DataLab itself is an isolated environment within the Metal Cloud and even within the same client infrastructure.

Data from existing databases such data warehouses or production systems can be imported securely into the DataLake via the VPN.

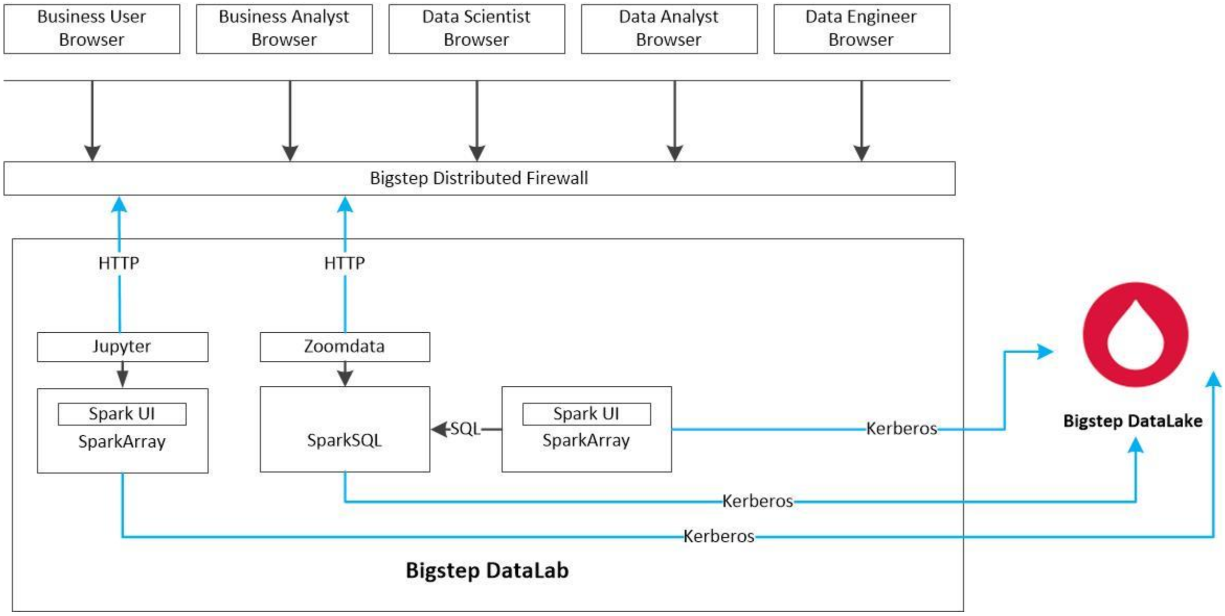

Figure 6: Bigstep DataLab Security Components

Performance

Bigstep uses state-of-the-art in-memory processing, baremetal nodes and direct connectivity.

The Spark cluster is designed to run a specific notebook at a certain point in time, focusing on delivering an environment used for data-exploration scenarios. Since customers can run multiple Spark clusters on the same infrastructure, Bigstep recommends using smaller Spark clusters for specific tasks instead of one big cluster that has to be shared among multiple users and multiple workloads. The goal is to have an easier way to fine-tune the memory requirements for a particular Spark job.

Using environment variables and the Spark API, available in the Jupyter Notebook, users can further fine-tune their Spark cluster.

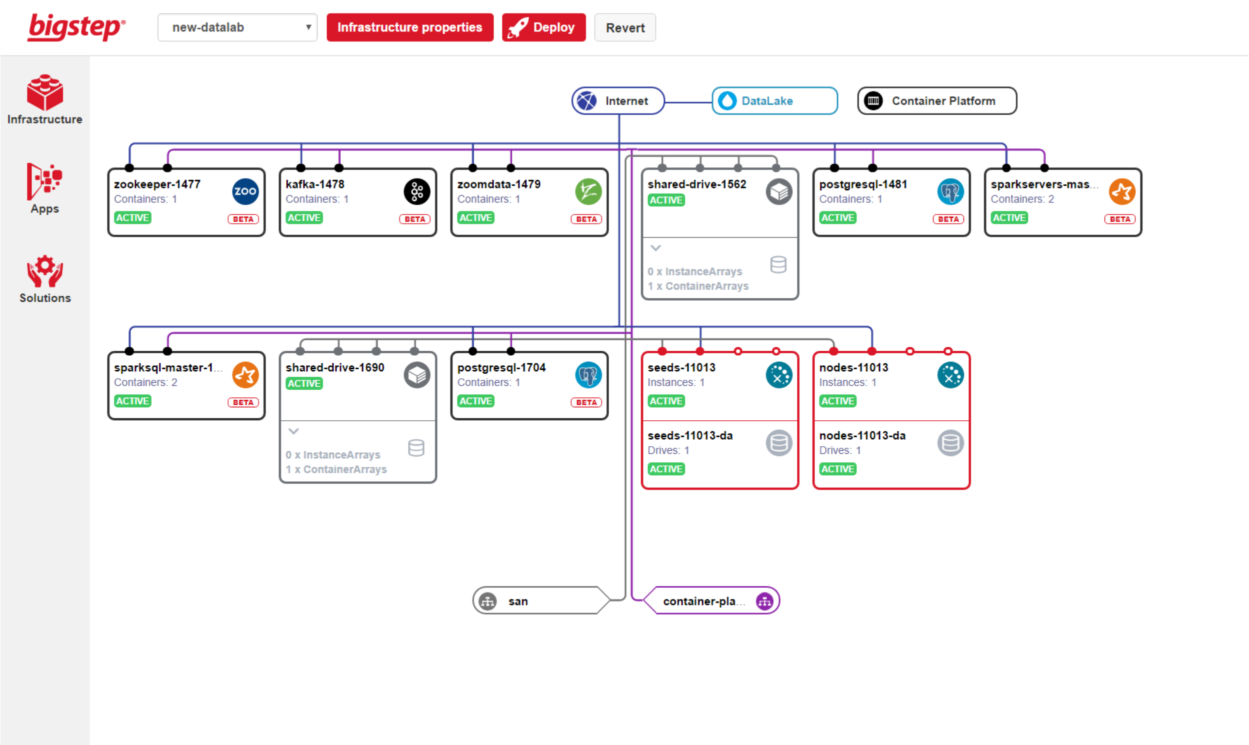

Figure 7: Deploying Bigstep DataLab on the Bigstep Platform

About Bigstep

Bigstep enables organizations to effortlessly create, launch and scale robust, secure and cost-effective data processing pipelines. From the underlying bare-metal infrastructure to the most valuable big data technologies Bigstep provides companies with all the key components required to make sense of their data.

Bigstep serves a global customer base from multiple data centers in Europe and the U.S. and has dual headquarters in London and Chicago.

For more information, please visit bigstep.com.