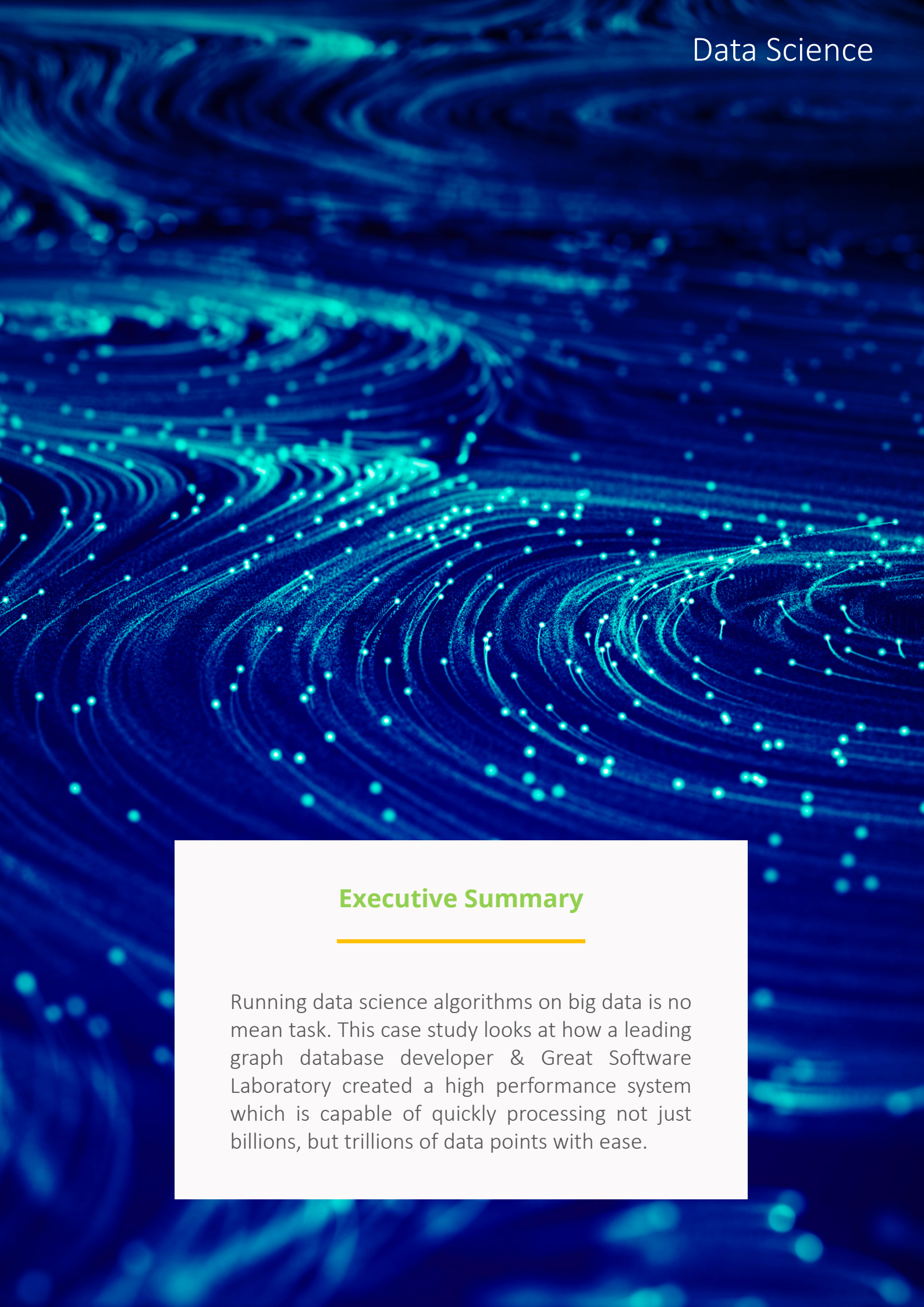
Data Science

Case Study

gslab

Empowering graph database users with high performance data science algorithms

Enterprise users get the ability to process massive data sets and do trillions of calculations easily using data science and ML



Executive Summary

Running data science algorithms on big data is no mean task. This case study looks at how a leading graph database developer & Great Software Laboratory created a high performance system which is capable of quickly processing not just billions, but trillions of data points with ease.

Overview

Our customer is a market leader in enterprise big data management and exploratory analytics. They help build large enterprise-wide data lakes which allow customers to consolidate and build data storage & retrieval facilities. This enables data management & analytics technologies to explore the data and represent it in powerful high resolution analytics & dashboarding solutions via their graph database product.

Our customer wanted to enable their users to run data science and feature engineering algorithms on the massive data inside their graph database product. This would empower customers to unearth powerful insights.

Challenge

Graph database users have massive enterprise data lakes. Enterprises use these to run various data science algorithms while finding patterns and insights. Features can also be extracted from these data lakes for ML model training. However, there were two challenges:

- Ready-to-use commercial libraries did not scale well for such computationally intensive processes. Existing libraries were not suitable for the product's distributed architecture that works with MapReduce methods. Because of these limitations:
    1. Analysis took a lot of time.
    2. It required a lot of computational resources.
    3. Libraries did not have flexibility to run on different infrastructures of choice.
    4. They were not scalable.

- The users of these algorithms are not necessarily data science experts. Hence there was a need to provide guidelines on using the right algorithms at the right times.

Solution

The graph database product exposes a number of extension points through which developers can customize and extend the system. The extension point interfaces and the user code that implements them are called User Defined eXtensions (UDXs). The APIs for these UDXs are available in C++ and JVM.

Great Software Laboratory developed more than 40 algorithms for data science & feature engineering as UDXs. These are an interesting prelude to machine learning and AI. All these algorithms are written in C++. The algorithms run in Massively Parallel Processing (MPP) clusters. Because of this architecture, they can work with different physical, virtual and cloud environments. This not only enables analysis of massive data but also offers superior analytical performance.

We included a broad variety of algorithms like distribution, feature exploration, sketches, correlation and profiling categories.

Empowering users

In order to help users in effectively using these algorithms, we worked on Zeppelin notebook. Each page:

- Demonstrated product implementation of the algorithm.
- Described the algorithm to users who were not data scientists.
- Provided easy examples for understanding.
- Provided a data set that allowed users to try out different algorithms.
- Provided SPARQL queries that demonstrated use of the algorithm.
- Provided details about the function of the SPARQL query.
- Explained the results and insights from the SPARQL queries.

Our solution covered the following algorithms:

Distribution: Normal, Poisson, Skellam, Continuous Uniform, Discrete Uniform, Student's t-Distribution, Weibull, Bernoulli, Binomial, Chi-squared, Exponential, Hypergeometric, Laplace, Log Normal, Logarithmic Series, Negative Binomial

Feature Exploration: Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Linear Discriminant Analysis (LDA)

Sketches: Cardinality Prominence Metric (HyperLogLog), Frequent Items, Theta Sketch (Set operations), KLL Sketch (Quantile/Rank)

Correlation: Pearson Correlation Coefficient, Matthews Correlation Coefficient, Spearman's Rank Correlation Coefficient

Profiling: T-Digest Metric, Geometric Mean, Skewness Metric, Discrete entropy

Impact

Improved performance

Empowered users

- There is a significant improvement in analytics and transformation performance using UDX even when it scales to billions or trillions of triples.
- Our solution empowers both data scientists and users who are not data scientists to generate powerful insights from massive enterprise data clusters.

Great Software Laboratory (GS Lab) has been the technology partner of choice to 100+ organizations across North America, Europe and Asia-Pacific for over 17 years. Leveraging our expertise in 130+ tools & technologies, we have created 300+ 'first-of-its-kind' solutions to real-world problems. Our 'Beyond code' philosophy ensures that we not only push boundaries of existing technologies but also try out newer problem solving approaches to keep our customers one step ahead of their competitors. Our global team of 1200+ employees is adept at creating 'real value' at each stage of the customer growth journey, right from proof-of-concepts to completely scaled up products. For more information about our solutions & offerings, please visit www.gslab.com