Big Data In-Database Analytics

Badillo Jose, Ozer Buse

ABSTRACT

The growth of volume and variety of data along with the velocity it is produced has increased the complexity of building predictive models with speed and accuracy. To address this problem, the organizations are migrating to in-database processing platforms which reduce the volume constraints of traditional analytic tools and allows the processing of larger datasets by pushing down the computation to the database systems itself. In this paper, we highlight the advances of in-database processing and analytics.

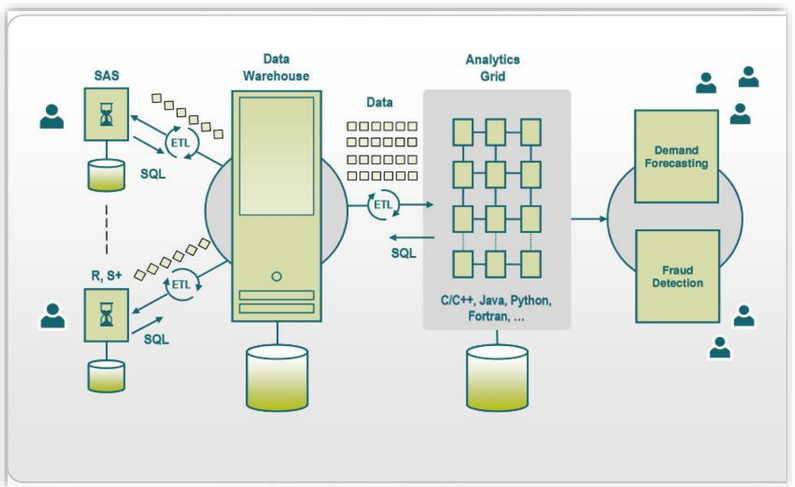
MOTIVATION

Traditional analytical tools first extract data from the database to working memory (In-application), to continue performing data preparation, integration, and cleaning tasks. Organizations are adopting in-database processing platforms to tackle the limitations introduced by the big volume and variety in traditional analytic tools. It allows the processing of larger datasets by pushing down the computation to the database systems itself.

In-database analytics, also known as in-database processing is a model where the data processing is performed into the database instead of shipping the data to analytical applications. Such a model aims to build analytical implementations in charge of computing statistics, correlations, feature selection, regression, or clustering analysis within the database where the data is located.

- Performance-enhancing features of the underlying database
- Eliminate the cost of transforming and loading data into other tools
- Reduces processing time from hours to seconds
- Avoids security breaches
- Guarantees the freshness of the data
- Accurate repeatable analysis

TRADITIONAL DATA ANALYTICS SYSTEMS



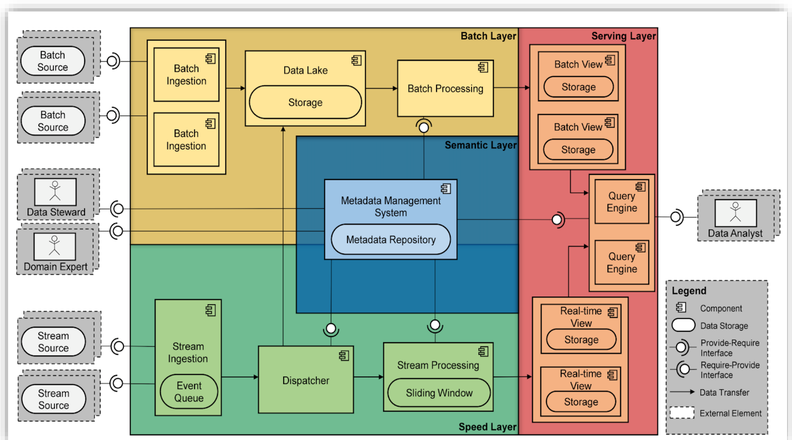
IN-DATABASE IN RELATIONAL DATABASES

In order to speed up critical analytical workloads, in-database analytics were performed in in relational databases with help of stored procedures. Stored procedures reside in a central database, help reducing network traffic and improving query performance, then data retrieval and analysis are much faster and data is more secure because it does not move from the enterprise data warehouse.

Categories	Function	Oracle Database	IBM DB2
Supervised Algorithms	Attribute Importance	Minimum Description Length	
	Classification	Decision Tree, Naive Bayes,	Decision Tree, KNN, Naive Bayes, Regression Tree
	Regression	Generalized Linear Models, Support Vector Machine	Linear Regression, Generalized Linear Models
Unsupervised Algorithms	Association	Apriori	Apriori, FT-Growth
	Clustering	K-Means, Orthogonal Partitioning Clustering	K-Means, TwoStep Clustering
	Feature Extraction	Non-Negative Matrix Factorization	
	Anomaly Detection	One Class Support Vector Machine	
	Dimension Reduction		Principal Component Analysis

BIG DATA ANALYTICS SYSTEMS

The growth of volume and variety of data along with the velocity it is produced has increased the complexity of building predictive models with speed and accuracy. To address this problem, the organizations are migrating to in-database processing platforms which reduce the volume constraints of traditional analytic tools and allows the processing of larger datasets by pushing down the computation to the database systems itself. In this paper, we highlight the advances of in-database processing and analytics.



DISTRIBUTED ML LIBRARIES

1-) ScalaNLP

Breeze is a numerical processing and machine learning.

Epic is a natural language processing and structured prediction toolkit written in Scala.(tokenization, sentence segmentation, syntactic parsing, named entity recognition).
Puck is a lightning-fast syntactic parser that uses GPUs to do its processing. Currently, it is only available in English.

2-) Spark Mllib

- Basic statistics
- Collaborative filtering
- k-means Clustering
- Dimensionality reduction
- Feature extraction and transformation
- Optimization
- Classification and regression

Problem type	Supported methods
Binary Classification	linear SVMs, logistic regression, decision trees, naive Bayes
Multiclass Classification	decision trees, naive Bayes
Regression	linear least squares, Lasso, ridge regression, decision trees

3-) Flink MLlib

- Machine Learning suitable for stream processing
- Iterates its data by using its streaming architecture

Problem type	Supported methods
Pipelines of transformers and learners	
Data pre-processing	Feature scaling, Polynomial feature base mapper
Supervised learning	Optimization framework with Stochastic Gradient Descent
Generalized Linear Models	Multiple linear regression, Support Vector Machines
Recommendation	ALS

CONCLUSION

Nowadays most of the data required to build more accurate analytic models do not fit in memory. In consequence, parallelization, in-database analytics, and data localization have gained terrain and we can find several research groups and organizations in the Industry working on implementations related to this area. Unfortunately designing distributed ML algorithms is highly complicated and developing distributed ML packages becomes platform dependent. Furthermore, being an immature field we observed there are no standardized measures to evaluate distributed algorithms. In the other hand, if we would like to go deeper into in-database we should look for processing in lower layers of the architectures. One of the approaches that could be implemented is pushing operations to disk. Acknowledging, that filtering operations happen very late in most cases when a query is executed, the aim of pushing predicates to disk is that moving filtering to an earlier phase boosts the query execution performance.