ABSTRACT

Applying Big Data technologies to high tech manufacturing

Dr. Dirk Ortloff, Nils Knoblauch

camLine GmbH

Presenter: Dr. Dirk Ortloff

Motivation

The systematic analysis of ever-increasing data collection presents companies with ever-greater challenges. Many companies simply lack the know-how to handle big data projects. Following to the motto "Let's do a Data Lake first", they bring together all available data in one system. Because often they are subject to the misconception that you should put as much data as possible in the system in order to gain the maximum insight and the most flexible evaluations. Unfortunately this does not work, because we can expect performance problems here. Therefore, the company must also think about meaningful evaluations for big data analytics in advance, that offer added value considering the cost-benefit ratio.

Description

Within the research and development project PRO-OPT, which is funded within the framework of the technology program "Smart Data - Innovationen aus Daten" of the Federal Ministry for Economic Affairs and Energy of Germany, camLine has taken on this problem. In collaboration with Audi, Continental, DSA, DFKI and Fraunhofer IESE the PRO-OPT project results enable companies in decentralized cooperative structures (smart ecosystems) to effectively and intelligently analyze large amounts of data. Digitalization and automation in particular are generating ever larger amounts of data in production. The data sources are distributed at different locations of a participant or at different, economically independent participants of the ecosystem. Overarching analysis must be broken down, taking into account access permissions to these sources. Big Data strategies should help to make these analyses more efficient.

As part of the PRO-OPT project, a wide variety of production data modeling approaches of were tested out. Apart from the problems of systematically merging different data buckets and the possible modeling of the data in NoSQL databases, the main focus of the work was on the analysis of these large data collections. The objective was to be able to apply and compare statistically reliable analyses and classification procedures as well as new procedures from the upcoming AI instruments.

To quickly gather experience in analyzing very large amounts of data and the pros and cons of NoSQL databases, the camLine product Cornerstone was used to work with Apache Cassandra. Two different paths were taken. On the one hand, the Cassandra database was accessed directly from Cornerstone and, on the other hand, via the tool stack shown in Figure 1.

ABSTRACT

Innovation

The innovation of the approach taken is twofold. The first important step is to use an analysis enabling data modeling approach. To easily and efficiently analyze data and find new correlations and root-causes, it is required to be able to store all data of a manufacturing step together with the corresponding metrology results along the whole manufacturing flow in a single row. Typical SQL database only allow rows to have up to 1000 columns. This amount of columns is not appropriate for the described merge of data in one row. To address this, Apache Cassandra was introduced and different data modeling approaches with up to 200.000 columns were used. Because datasets could reach several 10th of Terabytes, traditional statistical data analysis tools are not capable to cope with these volumes. The second innovation of the approach taken in PRO-OPT is the combination of traditional statistical methods with new Big Data analysis techniques applied to high tech manufacturing from a single environment, camLine Cornerstone. Using R through camLine Cornerstone gave the opportunity to have one front end working with Big Data and with small data. Applying the tool stack of Figure 1 huge amounts of data in the Cassandra cluster can be analyzed using Spark-R and the diverse analysis capabilities of R on a big multi-node computer cluster. Being able to “brush through R” iteratively zeroing in on a certain suspicious data cluster of anomalies was made possible. On a sub-selected, suitable size data set, the normal statistical analysis concepts of Cornerstone can be applied as presented in Figure 4. This way a lot of analysis flexibility for various data amounts were gained and even different approaches could be compared as shown in Figure 2.

Furthermore a new visualization technique was invented in the PRO-OPT project specifically suited for domains with high amounts of categorical data like semiconductor, photovoltaics and such. This so called Multi-Category-Chart is presented in Figure 3 and will be introduced in more detail during the presentation. Key point of the visualization is that it can be used to analyze up to 100 categorical variables at the same time in conjunction with brushing.

Results

This talk will show how the combination of the statistical data analysis system Cornerstone in conjunction with Apache Spark and Apache Cassandra provides a good basis for engineering analytics of massive data mounts. By properly nesting the solid mathematical methods in Cornerstone with big data-appropriate infrastructure such as Apache Spark and, in our case, Apache Cassandra, many new analytics issues can be addressed. Analyzes that used to be inefficient due to the sheer volume of data in classically modeled schema's can now be performed through appropriate big-table modeling and provide the ability to provide completely new insights into production data.

ABSTRACT

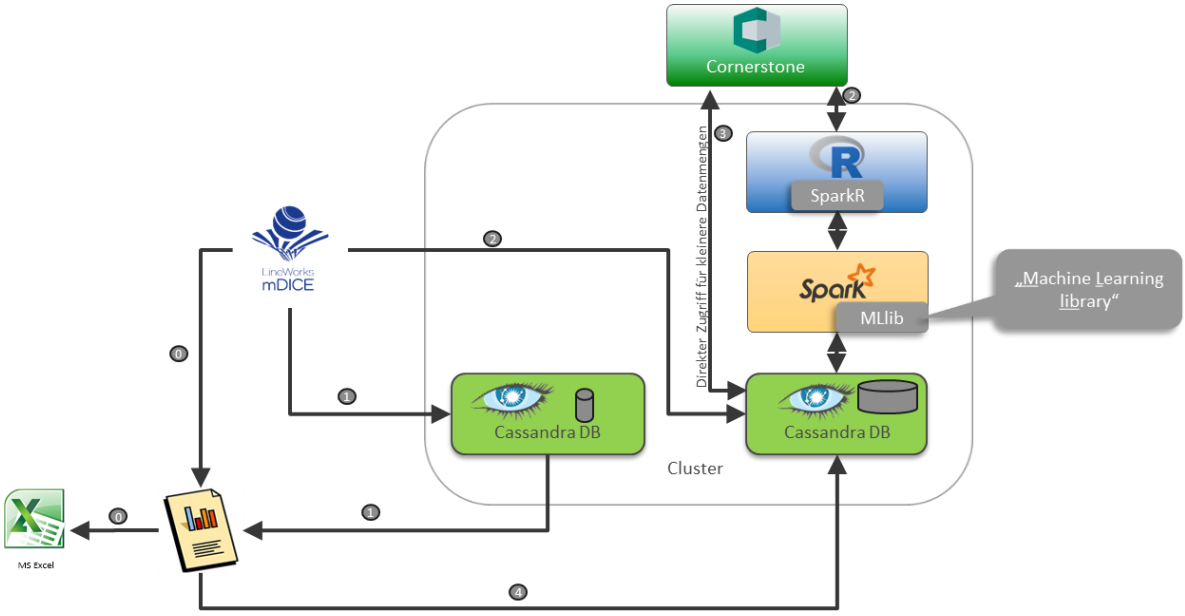

Figure 1: Tool stack

	Regression	Naive Bayes	Random Forest
Categorical target values	only binomial	Multinomial	Multinomial
Continuous target values	Yes	No	Yes
Categorical influences	Yes	Yes	Yes
Continuous influences	Yes	No	Yes
Non linear influences of continuous factors	Currently on 2-factor interdependencies with the „:“-Operator		Yes
Identification important influences	~	+	++
prediction accuracy	+	~	~

Figure 2: Comparison of the classification methods

ABSTRACT

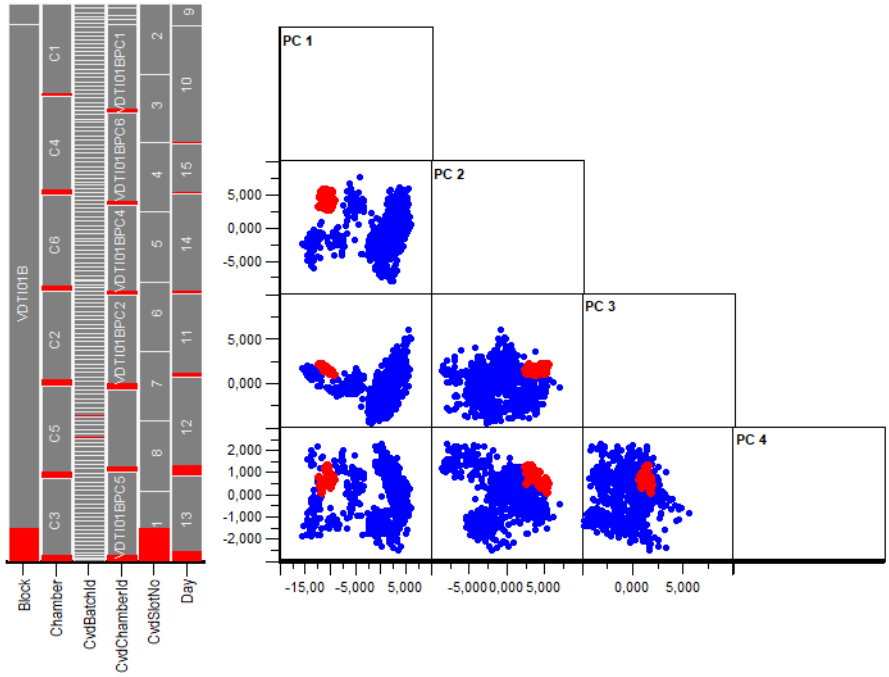

Figure 3: Example Multi-Category-Chart together with the brushed scatter plot

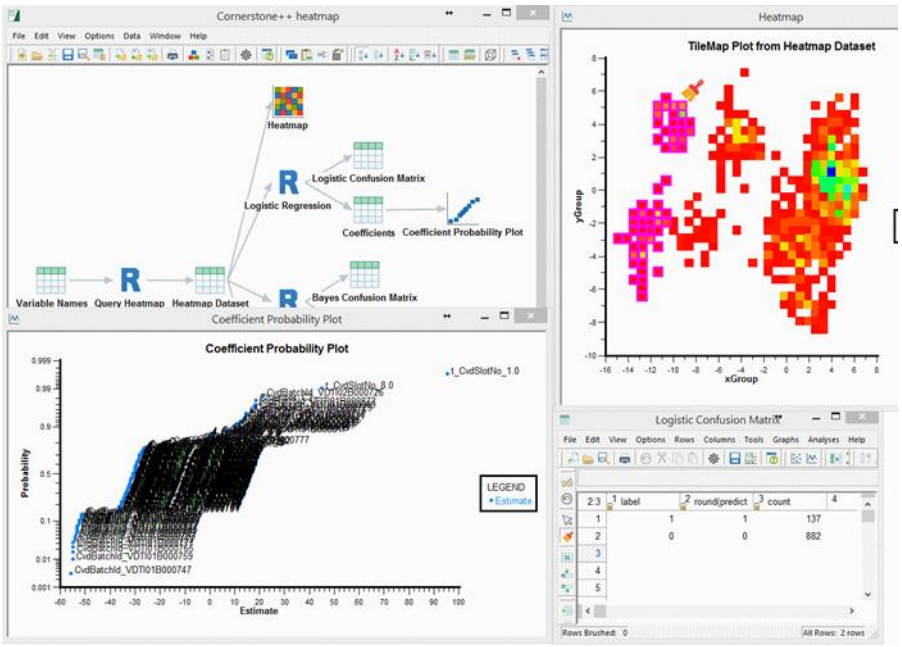

Figure 4: Example analysis results