Self-Service tools in the Big Data universe

Self-Service tools get more and more popular in the Big Data universe as many vendors are more focusing on not-so-tech-savvy business or analytics end users instead of the IT departments. Their motivation is to empower these stakeholders and facilitate the interaction between them and the IT. The longest existing products are Self-Service tools for Business Intelligence as visual analytics applications like *Tableau*. The visual interaction with the data enables end users to identify bottlenecks and derive hypothesizes and queries from the data as for example described by Dr. Matthias Weidlich for the clinical pathway analysis at the Dana Farber hospital.

Besides that there exist other interactive solutions for business users and analysts today that involve them at almost every process step from raw data to visualizations (fig. 1). Especially Self-Service Data Preparation tools help non-tech-savvy end-users to structure, join and explore their data sets and bypassing their yet biggest blocker in the Big Data process, iteratively cleaning and organizing data, which takes 60-80% of the time of Data Scientists as also mentioned in the intro event. What all these Self-Service applications have in common is that they are visual, interactive, easy-to-use and often predictive applications that give end users the ability to speed up their decisions. This poster will depict technology trends for end users with a selection of standalone-tools and later give a short outlook to the future of Self-Service tools in the Big Data jungle.

Technology Trends …

… in the field of Data preparation

Self-Service Data Preparation tools like the ones from *Trifacta*, *Paxata* or *Alteryx*, leaders in the field of Data Wrangling, are visual interactive applications that help data analysts from the extraction and curation to the explorative analysis of structured and unstructured data.

Therefore the trend are proprietary Machine Learning algorithms in combination with NLP like intelligent indexing, textual pattern recognition, and statistical graph analysis. In case of *Paxata* Machine Learning algorithms parse structured and unstructured data and build a flexible and comprehensive data model in form of a graph which represents relationships between data items. Knowing similar data items is then the basis for giving the users recommendations for data set joining. Furthermore syntactic and semantic quality issues are detected through the associations between the data. Besides that reinforcement learning algorithms improve usability from user interactions.

While *Paxata* always uses a Spark cluster as engine the *Trifacta* products automatically decide which engine to use depending on the data size and transformation operation (fig. 2). For smaller data sizes that only require a single node the preferred solution is their own In-Memory engine called Photon, for bigger datasets *Trifacta's* intelligent execution architecture runs the tasks on either Spark, MapReduce or Dataflow clusters (figure 2).

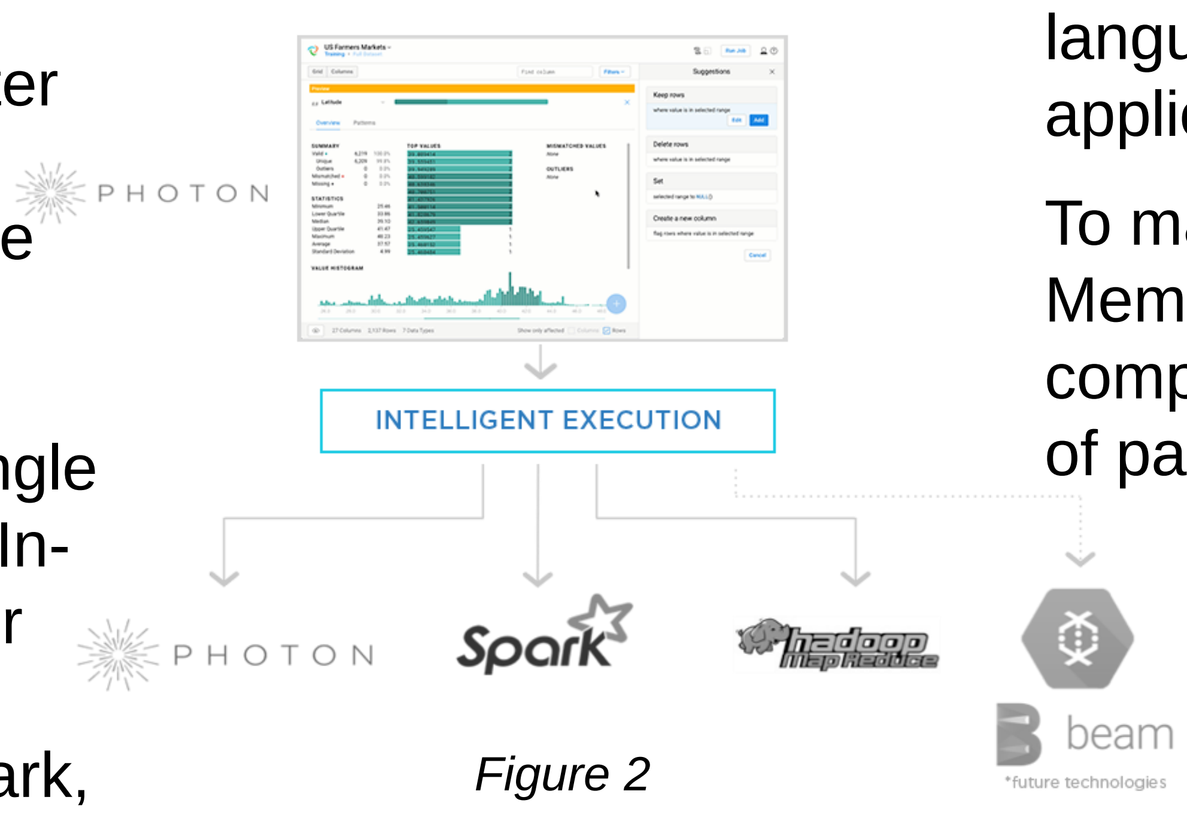

Figure 2

Besides the data wrangling products another important aspect for end users is data governance and transparency. For that purpose visual and interactive tools like *Alation* help users with the discovery and structuring of their data. Therefore the tools crawl metadata from diverse data lakes and other sources of an enterprise and hence index it in smart data catalogues. In a visual interface users can then explore their data set correlations with natural language search while Machine Learning and NLP give query predictions based on usage of data, data profiles and data lineage from query logs.

Fortunately for end users data catalogue tools like *Alation* can be combined with graphical data preparation tools like *Paxata* and *Trifacta* which then enables them "Self Service Data Discovery and Preparation" by easily importing and exporting specific data sets or data joins from one tool to the other. Also all of the mentioned tools offer interfaces to *Tableau* and other Self Service BI applications and hence make data exploration and analysis much easier for end users.

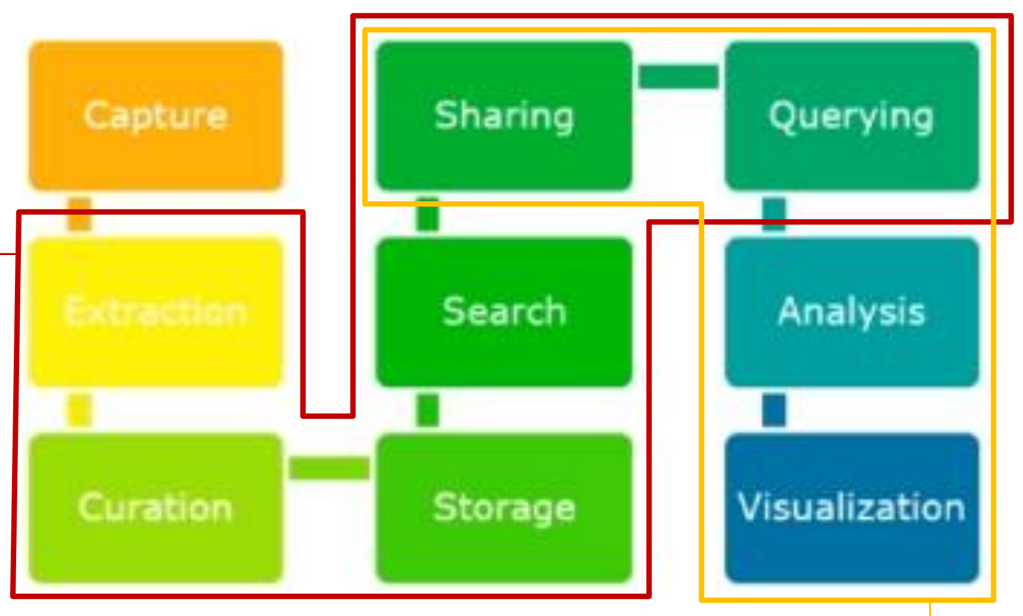

Figure 1

… in the field of Visual Analytics

Longer established Self-Service tools are visual analytics applications that replaced traditional BI tools. The leaders in the BI Magic Quadrant, *Tableau*, *Microsoft*, *Qlik Sense* and also *SAP Lumira* once again convince with easy-to-use, highly interactive and performant applications to analyze scaled-down data.

How Tableau works

As described by Lennart Heuckendorf the core of *Tableau* is VizQL, a query language that translates queries from visual interfaces like the *Tableau* application to SQL and back.

To make the application live-interactive *Tableau* integrated their new In-Memory Data Engine called Hyper in January 2018 that just-in-time compiles to and executes queries in Low Level virtual machine code instead of parsing it to C programs first.

To improve processor usage and reduce feedback time one main goal of Hyper was to achieve the best-possible parallelization in multi core systems. Therefore it splits tasks into small so-called morsels as depicted in figure 3.

The smaller segmentation gives Hyper more flexibility to handle requests with different priorities and respond to differences in core speed which result in better hardware utilization and faster performance.

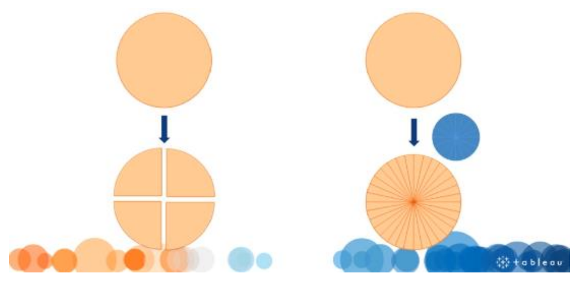
Figure 3

Future Vision

According to the *International Data Corporation* the amount spent on self-service visual discovery and data preparation tools will grow 2.5x faster than traditional IT-controlled tools for similar functionality like ETL-tools until 2020. Nevertheless the need for a thorough ETL-process remains as data quality needs to be assured and data needs to be stored in centralized enterprise warehouses for increased governance, reporting and other applications.

Furthermore according to Paul Boat from *Amitech* solutions there could be a "shift in the process of identifying relationships between data rather than the procedural steps required to transform data from one format to another" because current tools "start from where the data is rather than where the users are". The possible change would then imply starting at the Business Model and working backwards towards the data source.

Julius Rudolph

Master Student, IT-Systems Engineering
Hasso Plattner Institute, Potsdam, Germany

E-Mail: julius.rudolph@student.hpi.de