Title: Using big data in plant breeding

Academic team leader(s): Prof. Marcel Reinders (TU Delft)

Involved partners: Virtual Lab for Plant Breeding (Bejo Zaden, Bayer CropScience, Rijk Zwaan)

Challenge:
Currently, data is generated for thousands of accessions with each containing tens of millions of markers. Having many (e.g. over 1000) high-density genotyped individuals available for per species will become reality for breeding companies within the coming years. The de facto standard for storing variants is in specialized compressed binary files (BCF/VCF), which are indexed to allow for random access to specific genomic positions. Although these positional queries are very fast, they do not allow for flexible and fast interrogation of the data on other features than position.

Both in the academic and commercial arenas solutions are arising targeting the above from different angles. Here, some examples of public and private efforts are described, although this list is by no means exhaustive. Big Data Genomics is developing a genomics API on top of a stack of existing big data solutions as Avro, Spark and Hadoop. Google Genomics is developing an API to process and analyze genomic data allowing the use of their BigQuery infrastructure to explore genetic variants. Google offers the genetic variants resulting from the human 1000 genomes project preloaded in their database. In addition to the above commodity hardware systems, solutions running on specialized hardware are being developed. For example, the company Sqream has a GPU-backed database system with a genomics API called GenomeStack. The TU Delft spin-off Bluebee genomics is developing FPGA-accelerated versions of existing bioinformatics algorithms.

Subdisciplines relevant for this challenge

- Distributed systems
- Information management
- Bioinformatics / genotyping

The solutions space is large. A plethora of infrastructure and analytics solutions is available for generic applications in Big Data (Figure 1). It is not obvious which solutions are most suitable for genomics applications and form robust solutions for breeding companies towards the future. The objective of this project is to investigate possible Big Data solutions for high performance and flexible querying, computation and analysis on billions of genotype scores. For applications of genomics big data in breeding companies can we simply pick an existing solution and implement it? Do we need to pick several components from the large number of existing solutions and combine these in a way that suits our needs? Or do we have such specific and distinct requirements that we need custom built solutions?

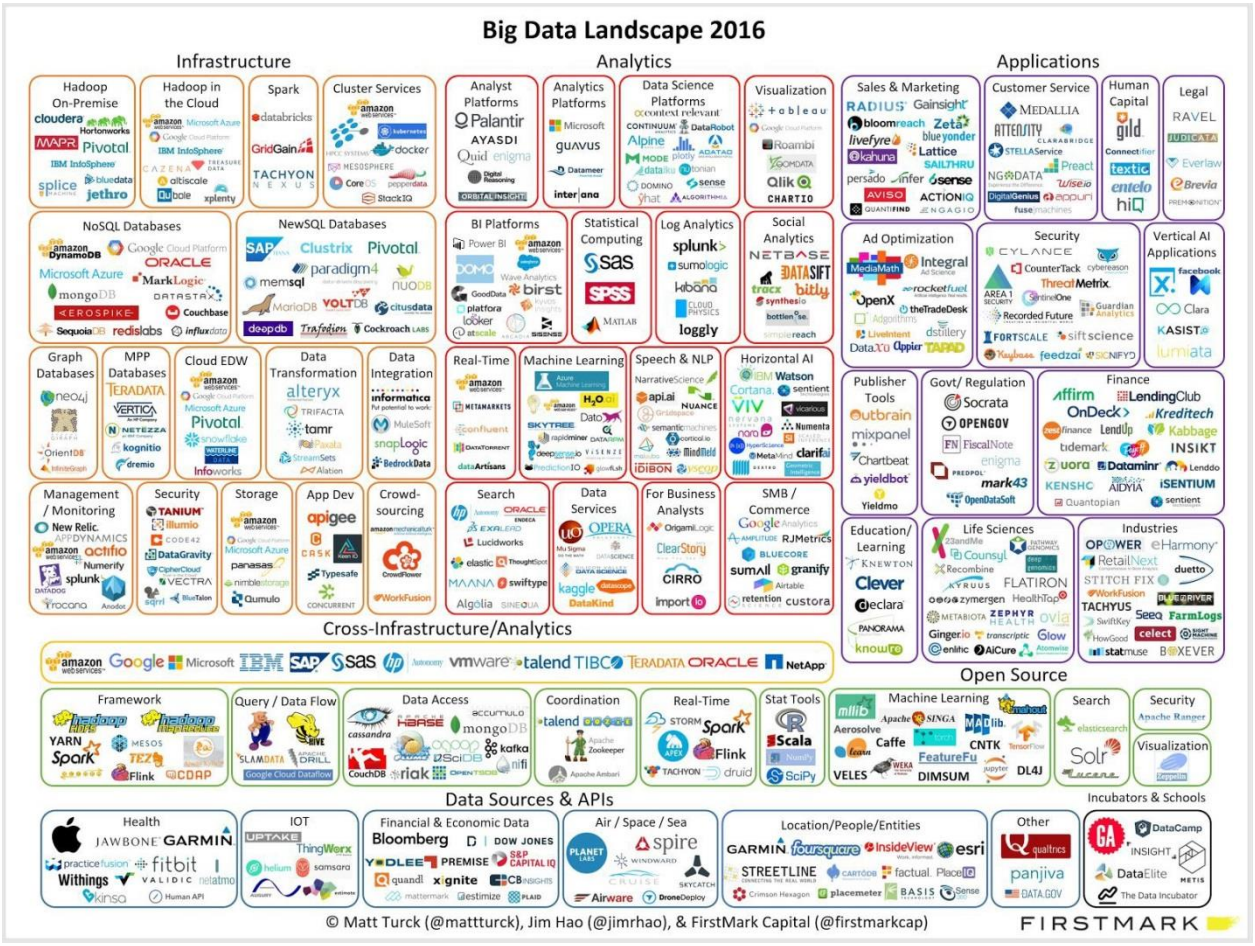

Figure 1 Big Data Landscape 2016 by Matt Turck

Boundary conditions

- We only consider genotype data and accompanying metadata. However, possibility to update (versioning) the physical location of the marker as the assembly improves, is part of this project.
- Participating companies have different requirements on data security (e.g. on premise vs. off-site).