INDUSTRY BRIEF

Whole Genome Research Drives Healthcare to Hadoop

Introduction

Where the cost and difficulty of sequencing the complete genome had historically been a barrier, advances in big data management, storage, and processing—especially the popularity of Apache Hadoop—are driving whole genome analytics towards the mainstream for research hospitals, biotech, and pharma. Whether the vector is breadth of global national projects or insertion into a wide variety of study types or presence in non-human, non-drug use cases or standardization as part of clinical genetics, whole genome has become a major theme in the health and life science domain.

Whole genome has resulted in rapid adoption of next-generation sequencing applications, replacing custom and exome sequencing in many cases, and generating huge data sizes. Just as the plummeting cost of whole genome sequencing has contributed to its early arrival, big data technology has also created a commanding set of analytics options for researchers, clinicians, principle investigators, and forensic geneticists.

> "Our clients are reporting that the [Cloudera] system has actually saved hundreds of lives by being able to predict if a patient is septic more effectively than they could before."
>
> Cerner

A Full Spectrum of Use Cases and Opportunity

Today, national governments, pharmaceutical companies, health systems, and research organizations use whole genome data. There is no part of the spectrum of health and life science practitioners that will not soon leverage the full DNA sequence. In the future, food and consumer goods manufacturers, law enforcement, state and civic organizations, and altogether unexpected commercial groups will also join in building whole genome capabilities, particularly as technology becomes more accessible and affordable beyond North America and Europe.

Industry	Use Case for Whole Genome Sequencing Analytics
Research Hospitals	St. Jude's Children's Research Hospital looks for pediatric cancer gene mutations across the spectrum of childhood cancers.[1]
Academia	Seoul National University, University of Arizona, and Kansas State University analyzed the whole genome of the Asian honeybee for the first time and identified differences in odor perception from the western honeybee.[2]
Biotech	Gilead Sciences uses whole genome sequencing to analyze tumor cells.[3]
Payers	While adoption is limited today, there have been cases of insurers agreeing to pay for whole genome sequencing for diagnostic purposes.[4,5]
Government	Analysis found that each additional year of a father's age when his child is born increases the likelihood of gene mutations of the child by 2.5%.[6]
Consumer Goods	Procter & Gamble has sequenced the fungus that causes dandruff, ideally leading to insights about how to affect causative factors.[7]
Food	The U.S. Food and Drug Administration proved a national outbreak of *Listeria monocytogenes* originated from a specific food manufacturer.[8]
Intelligence and Law Enforcement	The FBI completed an analysis of powder sent to various addresses in 2001 to determine contamination with anthrax and other dangerous substances.[9]
Manufacturing	ExxonMobil analyzes the toxicity of substances in its plasticizers.[10]

[1] St. Jude's Children's Research Hospital and Washington University in St. Louis School of Medicine. "Identifying Cancer Mutations." PediatricCancerGenomeProject.org.

[2] Park, Doori, et al. "Uncovering the Novel Characteristics of Asian Honey Bee, *Apis cerana*, by Whole Genome Sequencing." *BMC Genomics*. 2 January 2015.

[3] Leuty, Ron. "Gilead Cancer Strategy Relies Heavily on Yale." San Francisco Business Times BiotechSF blog. 2 December 2011.

[4] Herper, Matthew. "Sequencing a Child's DNA—And Getting an Insurance Company to Pay." Forbes.com. 2 March 2011.

[5] MacDougall, Raymond. "New Report Offers a Primer for Doctors' Use of Clinical Genome and Exome Sequencing." NIH.org. 18 June 2014.

[6] Koboldt, Dan. "Population Whole-Genome Sequencing: Dutch Edition." MassGenomics.org. 4 August 2014.

[7] Dawson, Thomas. "Malassezia globosa and restricta: Breakthrough Understanding of the Etiology and Treatment of Dandruff and Seborrheic Dermatitis through Whole-Genome Analysis." *Journal of Investigative Dermatology Symposium Proceedings*. 2007.

[8] Acheson, David. "FDA FSMA Facility Supervision Powers—Appropriate or Abusive?" AchesonGroup.com. 20 March 2014.

[9] National Research Council. *Review of the Scientific Approaches Used During the FBI's Investigation of the 2001 Anthrax Letters*. National Academies Press. 2011.

[10] Plummer, Simon M., et al. "Identification of Transcription Factors and Coactivators Affected by Dibutylphthalate Interactions in Fetal Rat Testes." *Toxicology Sciences*. 28 February 2013.

> “We are at the cutting edge of disease prevention and treatment, and the work that we will do together will reshape the landscape of our field. Mount Sinai is thrilled to join minds with Cloudera.”
>
> Mount Sinai School of Medicine

Given the availability of technology to analyze the whole genome, even greater opportunity lies in the combination with external data sets, many of which are public information. In fact, comparing internal data with public data sets is a requirement of any comprehensive analytic bench. Hadoop sits at the heart of the value that researchers, providers, payers, pharma, and biotech can gain from whole genome research because, as the core of an enterprise data hub, it not only scales to the massive data size required to host the sequence, but also combines this data with an enormous variety of descriptive and real-time data, all with full security, governance, analytics, and visualization options. The following are a set of general databases and browsers Cloudera most often encounters in this space.

Commonly Encountered General Databases and Browsers

Population Genomics	• 1000 Genomes • ExAC • dbSNP • Exome Variant Server
Clinical Genomics	• dbGaP Cohorts • OMIM • ClinVar • Disease-Specific Databases • HGMD
Functional Genomics	• Ensembl • ENCODE • UCSC Browser • GEO • UniProt/Swissprot
Cancer Genomics	• COSMIC • GTex • TCGA

Source: Cloudera

There is also a host of condition-specific, disease-related data sets, which can be even more valuable to a researcher. Better than any other data solution, including custom tools, an enterprise data hub scales to rapidly ingest and accommodate any source, size, or complexity of new data as research changes and grows over time.

Orders of Magnitude More Health Data

Beyond the breadth of data sources, industries, and research types, whole genome itself has led to an order of magnitude more data. It is now normative to find research hospitals using thousands of cases and controls, and biotechs, pharmas, and research consortia using tens of thousands of cases and controls.

How is the data volume landscape changing? Here are some recent figures from a research lab’s data statistics:

- Each study’s single whole genome pipeline produces 1x100GB BAM file, from which 4 million to 5 million variants are extracted into variant call file (.vcf) format.
- The per-sample requirement in storing data in a database ranges from 4 million to 3 billion table rows, depending on compression.
- Studies span across 10,000 to 100,000 controls and cases.
- Each study produces data in 40 billion to 300 trillion table rows.
- Processing requirements are scaling linearly.

To add context and a point of comparison, a single lab conducting a single whole genome study can produce more data than a very large retail chain might amass in its biggest database table during five or more of its busiest years. A lab with multiple research teams could have petabytes of active data to analyze, with an order of magnitude more requiring warm storage.

This challenge is not unique to large corporations or research consortia. This is a challenge affecting niche research hospitals, single study labs, and every type of whole genome research. There is a consistent ‘push towards *n*’—a desire to increase sample size, thereby decreasing statistical error, increasing statistical confidence, and making studies more robust and more likely to be published in the most prestigious journals. In other words, the business mission is more likely to succeed as the data pools of whole genomes increase in size, bringing in more cases and controls.

Integrating the Whole Genome Value Chain

With previous next-gen sequencing approaches, high performance computing (HPC) was a realistic and standard approach to running downstream analytic jobs. In an HPC platform, data is brought to compute by moving it into memory, enabling very fast calculations. As the nature and scope of data evolves, however, the time requirements for and cost of bringing such large data sets to compute make legacy approaches to HPC unfeasible and require new approaches to bringing compute to the data.

The current most common practice is to marry an existing HPC environment with a big data platform. Due to the maturity of upstream cleansing, quality control, alignment, and annotation tools written for HPC and network file systems, it is often easier to use existing tools for data pipeline and cold storage. However, those legacy architectures fail at performing analytics on such massive volumes of data. With custom sequencing or small and medium micro-arrays, a one-size-fits-all compute approach may function, depending on the complexity of the analytics.

Hadoop is designed to act as the hub feeding both cold storage and HPC at the appropriate points in the data value chain. There is little to no integration needed, because an enterprise data hub can store, process, and analyze data in any format.

Large technology, financial services, and telecommunications companies have faced a similar challenge, giving rise to widespread implementation of big data platforms in the commercial sector during the past decade. As part of an enterprise data hub, Hadoop can also help health and life science organizations analyze an entire genome in a reported hour or less (compared to weeks in the past), accompanied by a 99% accuracy rate at a cost in the hundreds of dollars—a thousand-fold improvement in expense and throughput. For whole genome, only Hadoop delivers the best architecture for the most relevant work and the greatest option value at lowest cost.

Customer Case Study	
Problem	With a NoSQL database, typical rare variant analytics on roughly 3,500 cases and controls took between four and seven days to complete for a single study. Most of that time was consumed moving data in and out of memory so HPC could act on it.
Solution	**Cloudera Enterprise: Data Hub Edition** • Apache Hadoop: an open-source software framework for storage and largescale processing of large data sets on clusters of industry standard hardware • Cloudera Manager: the first and most sophisticated management application for Hadoop and an enterprise data hub • Cloudera Navigator: the first fully integrated data security and governance application for Hadoop-based systems, providing full discoverability, lineage, and encryption with key management • Cloudera Impala: Hadoop's massively-parallel-processing SQL query engine • Apache Spark: the next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS via in-memory capabilities • Apache Parquet: a columnar storage format available to any project in the Hadoop ecosystem, regardless of processing framework, data model, or programming language • Apache Avro: a framework for serialization of persistent data and remote procedure call between Hadoop nodes and between client programs and Hadoop services
Impact	The time required to complete the same analytic workload decreased to under an hour. The bioinformatician was able to increase cases and controls to more than 10,000 per study while serving more research groups faster.

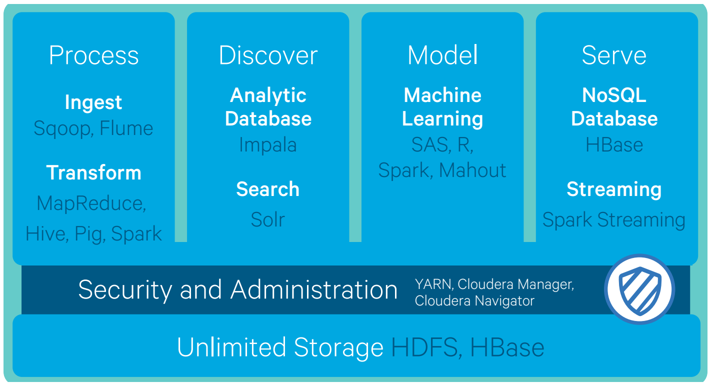
Source: Cloudera

Big Data and an Enterprise Data Hub

Healthcare is perhaps more affected by the rapid proliferation of data than any other industry. With HITECH and HIPAA regulations and high-visibility data breaches driving more focus on security, retention, and privacy of personal health information (PHI), the latent demand for a scalable data strategy has become a primary driver of industry decision-making. Rapid innovation, spearheaded by new technology and data-driven leadership, is upending the business of health. This brings analytics into the spotlight for a movement dedicated to efficiency, accessibility, and, most importantly, outcomes.

The past decade has seen an abundance of analytic offerings related to population health and medical home, disease-specific and cohort-specific outcomes, and the accountable care organization (ACO). Underlying all these new solutions is the challenge and opportunity of massive data growth. The potential payoff is huge: when collected, combined with public and historic information, and made secure, big data offers providers, payers, pharma, device manufacturers, and solutions developers deeper insights than previously imaginable.

Today, the introduction of an enterprise data hub built on Apache Hadoop at the core of your information architecture promotes the centralization of all data, in all formats, available to all users—analysts, clinicians, investigators, researchers—with full fidelity and security at up to 99% lower capital expenditure per terabyte compared to traditional data management technologies.



The enterprise data hub serves as a flexible repository to land all of an organization's unknown-value data, whether for compliance purposes, for advancement of processes core to the mission like whole genome sequencing, real-time biomonitor feed processing, sepsis prevention, and population cohort segmentation, or for more sophisticated applications such as machine learning models built on clinical text analytics with natural language processing.

It incomparably expands data sources and speeds up processing and analytics to deliver markedly better insights on the origin, nature, and proper treatment of many more diseases, toxins, and common ailments. And it increases the availability and accessibility of data for the activities that provide a full picture of operations at a health system, hospital, research organization, pharmaceutical company, or biotech firm to enable process innovation—all completely integrated with existing infrastructure and applications to extend the value of, rather than replace, past investments.

However, the greatest promise of information-driven health and life science resides in the questions that drive better outcomes that organizations have historically been unable or afraid to ask, whether because of a lack of coherency in their data or the prohibitively high cost of specialized tools. An enterprise data hub encourages more exploration and discovery with an eye towards helping decision-makers bring the future of their industries to the present:

> *Can we keep a decade of EMRs online to comply with HIPAA requirements and also make that data available for analysis alongside generic industry data and new patient data?*
>
> *How do we better prevent adverse effects using the massive trial data available on thousands of drugs, millions of compounds, and countless individual genetic variations?*
>
> *Can we personalize medicine by combining and analyzing clinical data on the whole genome sequence with public health databases and browsers?*
>
> *What is required to build predictive healthcare models that minimize emergency room visits by streaming patient data from remote, wearable sensors in real time?*

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Only Cloudera offers everything needed on a journey to an enter¬prise data hub, including software for business critical data challenges such as storage, access, management, analysis, security and search. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1400 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production. www.cloudera.com.