Data Lakes and Their Implication on the Global Health Sector v2.0

Research | Blogs | Expert Opinions

Pekka Neittaanmäki

Dean of the Faculty of Information Technology

Professor, Department of Mathematical Information Technology

University of Jyväskylä

Miikael Lehto

Project Researcher

University of Jyväskylä

Anthony Ogbechie

Service Innovation Management

University of Jyväskylä

JYVÄSKYLÄN YLIOPISTO

INFORMAATIOTEKNOLOGIAN TIEDEKUNTA

2017

TEKES HANKE: VALUE FROM PUBLIC HEALTH WITH COGNITIVE COMPUTING

TABLE OF CONTENTS

What is a Data Lake? 3

Data Lakes 101: An Overview 4

Data Lake 4

Data Lakes and the Promise of Unsiloed Data 6

Investing in a Data Lake? Shore Up the Big Data Gateway 7

Why Do I Need a Data Lake?..... 8

Data Lakes: The Biggest Big Data Challenges 9

Your 'Resolution List' for 2017: 5 Best Practices for Unleashing the Power of Your Data Lakes 9

Charting the Course Toward Value-Based Care With a Healthcare Data Lake 11

Diving in With a Healthcare Data Lake for Predictive Care 11

Why Healthcare Organisations Should Take the Plunge With Data Lakes 12

Partners Data Lake Offers Healthcare Analytics as a Service 13

Medical Insight Set to Flow From Semantic Data Lakes 13

What is a Data Lake?

A data lake is a shared data environment that comprises multiple repositories and capitalizes on big data technologies. It provides data to an organization for a variety of analytics processing including:

- discovery and exploration of data
- simple ad hoc analytics
- complex analysis for business decisions
- reporting
- real-time analytics

Organizations are increasingly exploring the data lake approach to address demands for an agile yet secure and well-governed data environment that supports both structured and unstructured data.

A data lake is:

- An environment where users can access vast amounts of raw data
- An environment for developing and proving an analytics model, and then moving it into production
- An analytics sandbox for exploring data to gain insight
- An enterprise-wide catalog that helps users find data and link business terms with technical metadata
- An environment for enabling reuse of data transformations and queries

A data lake is a storage repository that holds an enormous amount of raw or refined data in native format until it is accessed. The term data lake is usually associated with Hadoop-oriented object storage in which an organization's data is loaded into the Hadoop platform and then business analytics and data-mining tools are applied to the data where it resides on the Hadoop cluster.

However, data lakes can also be used effectively without incorporating Hadoop depending on the needs and goals of the organization. The term data lake is increasingly being used to describe any large data pool in which the schema and data requirements are not defined until the data is queried.

https://www.ibm.com/analytics/uk-en/data-lake/

Data Lakes 101: An Overview

A Data Lake is a pool of unstructured and structured data, stored as-is, without a specific purpose in mind, that can be "built on multiple technologies such as Hadoop, NoSQL, Amazon Simple Storage Service, a relational database, or various combinations thereof," according to a white paper called What is a Data Lake and Why Has it Become Popular?

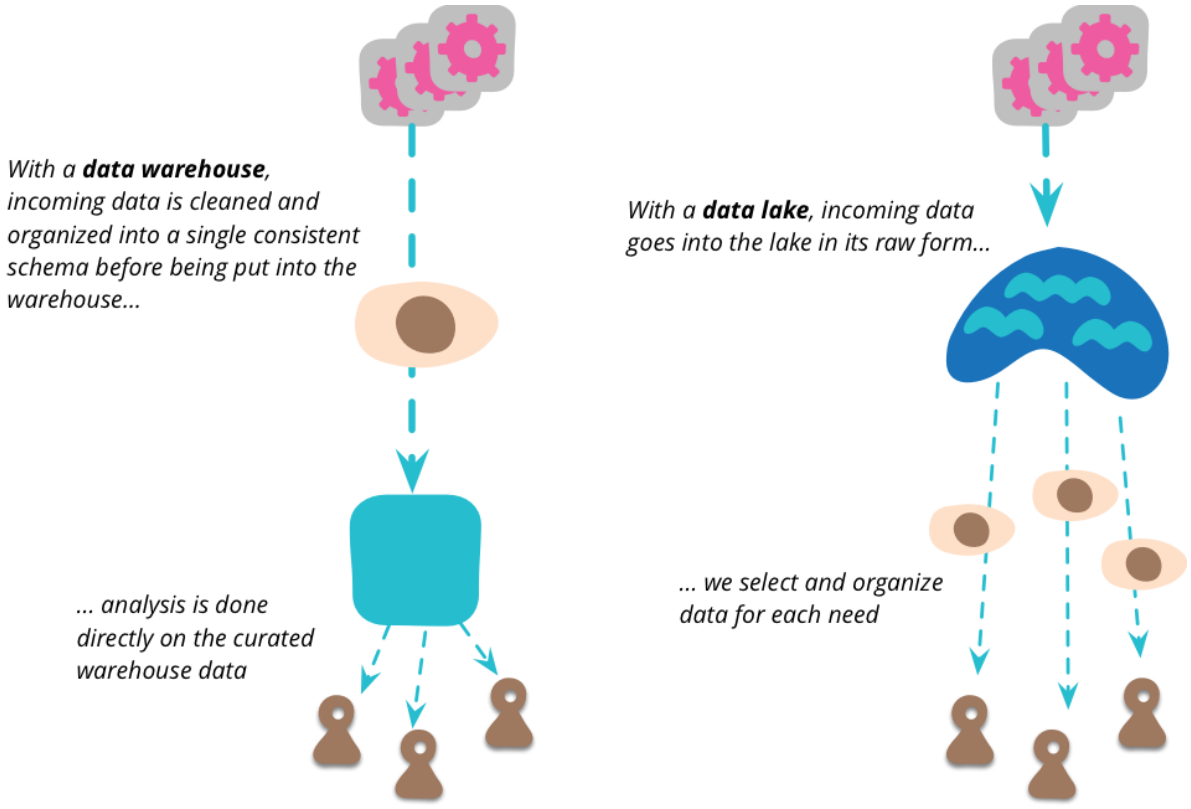
"A Data Lake is characterized by three key attributes:

1. **Collect everything**. A Data Lake contains all data, both raw sources over extended periods of time as well as any processed data.
2. **Dive in anywhere**. A Data Lake enables users across multiple business units to refine, explore and enrich data on their terms.
3. **Flexible access**. A Data Lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines."

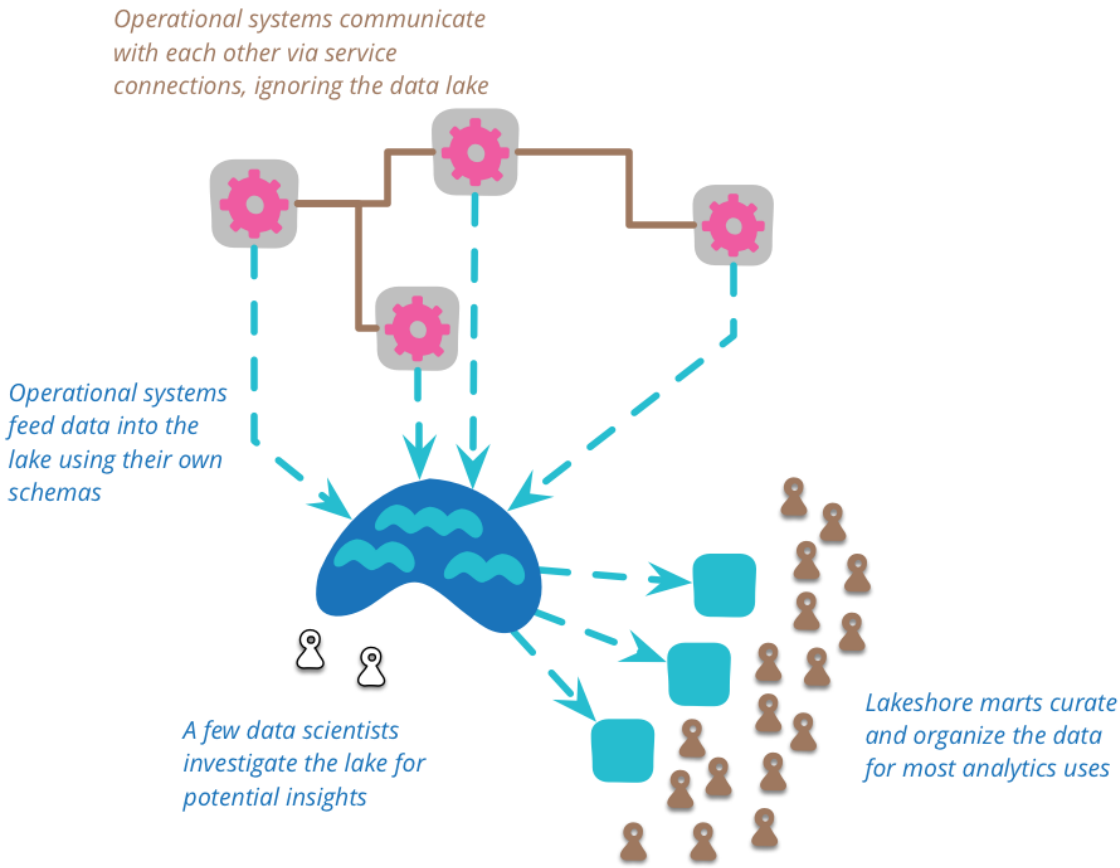
http://www.dataversity.net/data-lakes-101-overview/

Data Lake

There is a vital distinction between the data lake and the data warehouse. The data lake stores *raw* data, in whatever form the data source provides. There is no assumptions about the schema of the data, each data source can use whatever schema it likes. It's up to the consumers of that data to make sense of that data for their own purposes.

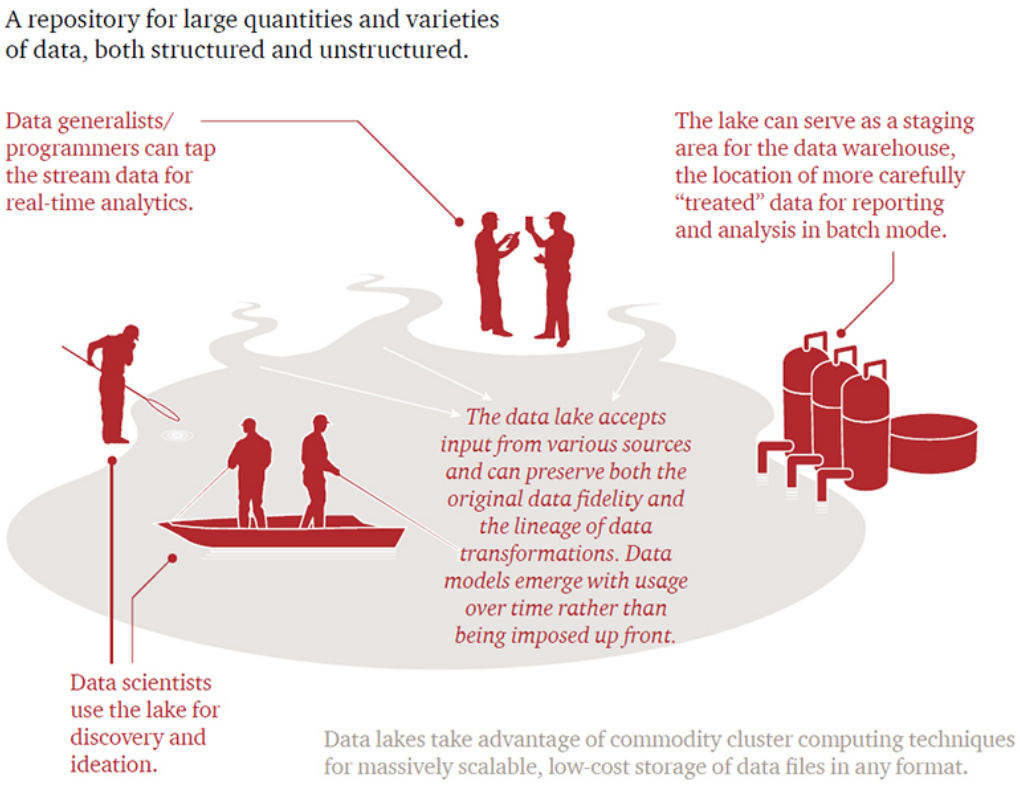


The complexity of this raw data means that there is room for something that curates the data into a more manageable structure (as well as reducing the considerable volume of data.) The data lake shouldn't be accessed directly very much. Because the data is raw, you need a lot of skill to make any sense of it. You have relatively few people who work in the data lake, as they uncover generally useful views of data in the lake, they can create a number of data marts each of which has a specific model for a single bounded context. A larger number of downstream users can then treat these lakeshore marts as an authoritative source for that context.



https://martinfowler.com/bliki/DataLake.html

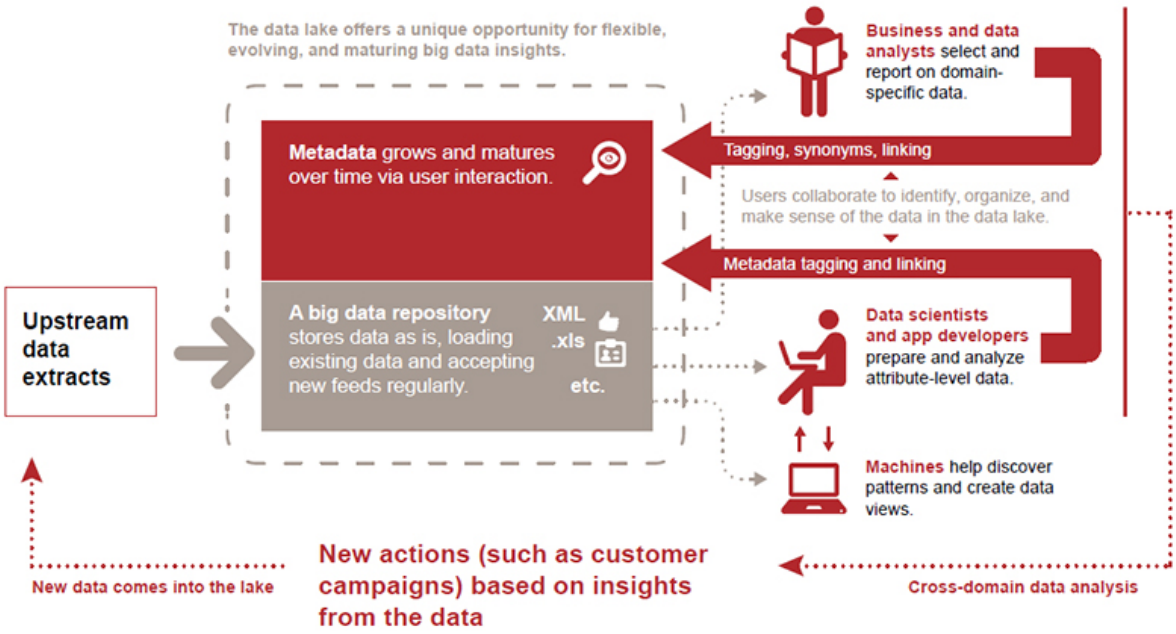
Data Lakes and the Promise of Unsiloed Data



Hadoop allows the hospital's disparate records to be stored in their native formats for later parsing, rather than forcing all-or-nothing integration up front as in a data warehousing scenario. Preserving the native format also helps maintain data provenance and fidelity, so different analyses can be performed using different contexts.

Previous approaches to broad-based data integration have forced all users into a common predetermined schema, or data model. Unlike this monolithic view of a single enterprise-wide data model, the data lake relaxes standardization and defers modeling, resulting in a nearly unlimited potential for operational insight and data discovery. As data volumes, data variety, and metadata richness grow, so does the benefit.

The data lake loads data extracts, irrespective of format, into a big data store. Metadata is decoupled from its underlying data and stored independently, enabling flexibility for multiple end-user perspectives and incrementally maturing semantics.



http://usblogs.pwc.com/emerging-technology/data-lakes-and-the-promise-of-unsiloed-data/

Investing in a Data Lake? Shore Up the Big Data Gateway

By capturing largely unstructured data for a low cost and storing various types of data in the same place, a data lake:

- Breaks down silos and routes information into one navigable structure. Data pours into the lake lives there until it is needed, when it flows back out again.
- Gathers all information into one place without first qualifying whether a piece of data is relevant or not.
- Enables analysts to easily explore new data relationships, unlocking latent value. Data distillation can be performed on demand based on business needs, allowing for identifying new patterns and relationships in existing data.

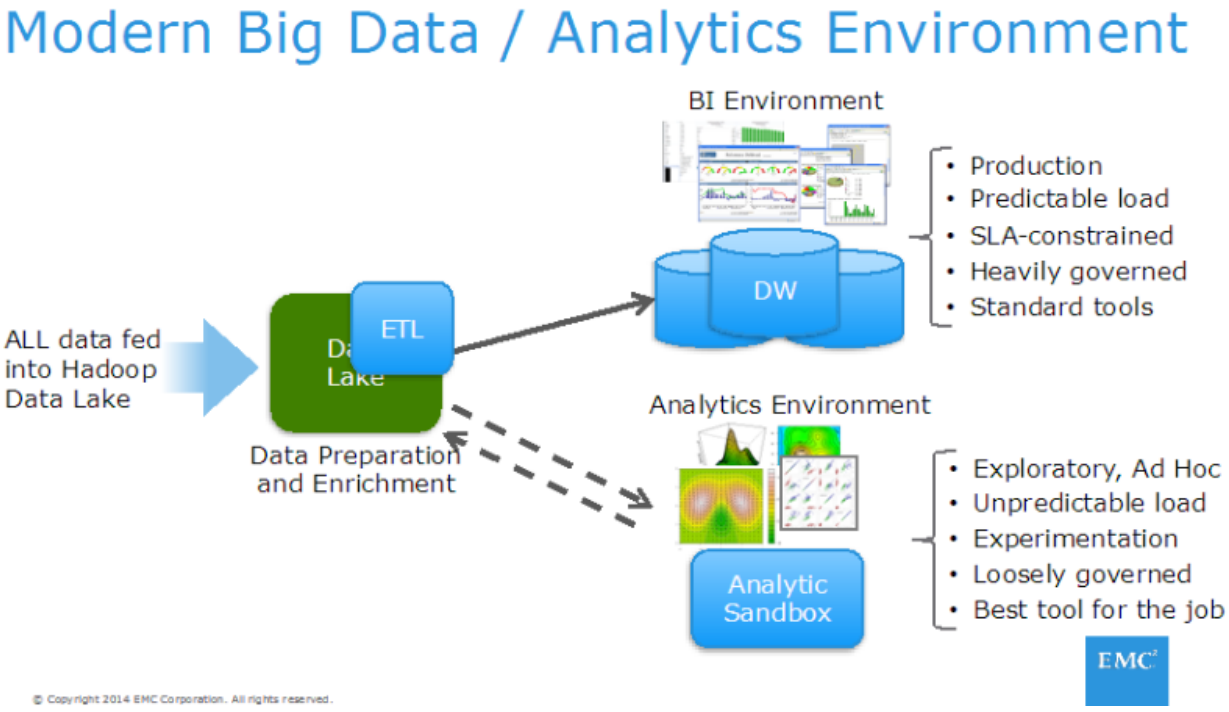
- Helps deliver results faster than a traditional data approach. Data lakes provide a platform to utilise heaps of information for business benefits in near real-time.

- Healthcare: Health systems maintain and analyse millions of records for millions of people to improve ambulatory care and patient outcomes. Quick insight and action on such records also can improve operational efficiency and enable an accountable care organisation.

http://www.itproportal.com/2015/09/25/investing-data-lake-shore-up-big-data-gateway-2/

Why Do I Need a Data Lake?

The data lake is a powerful data architecture that leverages the economics of big data (where it is 20x to 50x cheaper to store, manage and analyze data as compared to traditional data warehouse technologies). And new big data processing and analytics capabilities help organizations address business and operational challenges that were difficult to address using conventional Business Intelligence and data warehousing technologies.



https://www.linkedin.com/pulse/why-do-i-need-data-lake-bill-schmarzo

Data Lakes: The Biggest Big Data Challenges

Data lakes need to have four primary components – data ingestion, data management, query management and data lake management.

Data Ingestion

- Model driven
- Semantic tagging
- On-demand query
- Streaming
- Scheduled batch load
- Self service

Data Management

- Data movement
- Data provenance
- Types (In memory, NoSQL, MR, columnar, graph, Semantic, HDFS)
- Data flow (governed data, lightly governed data, ungoverned data)

Query Management

- Semantic search
- Data discovery
- Analytics directed to best query engine
- Capture and share analytics expertise
- Query data, metadata and provenance

Data Lake Management

- Models (Biz unit data optimized to assist analytics)
- Data assets catalog (ontologies, taxonomies)
- Workflow (processes, schedules, provenance capture)
- Access management (AAA, group/role/rule/user-based authorization)
- Metadata

http://analytics-magazine.org/data-lakes-biggest-big-data-challenges/

Your 'Resolution List' for 2017: 5 Best Practices for Unleashing the Power of Your Data Lakes

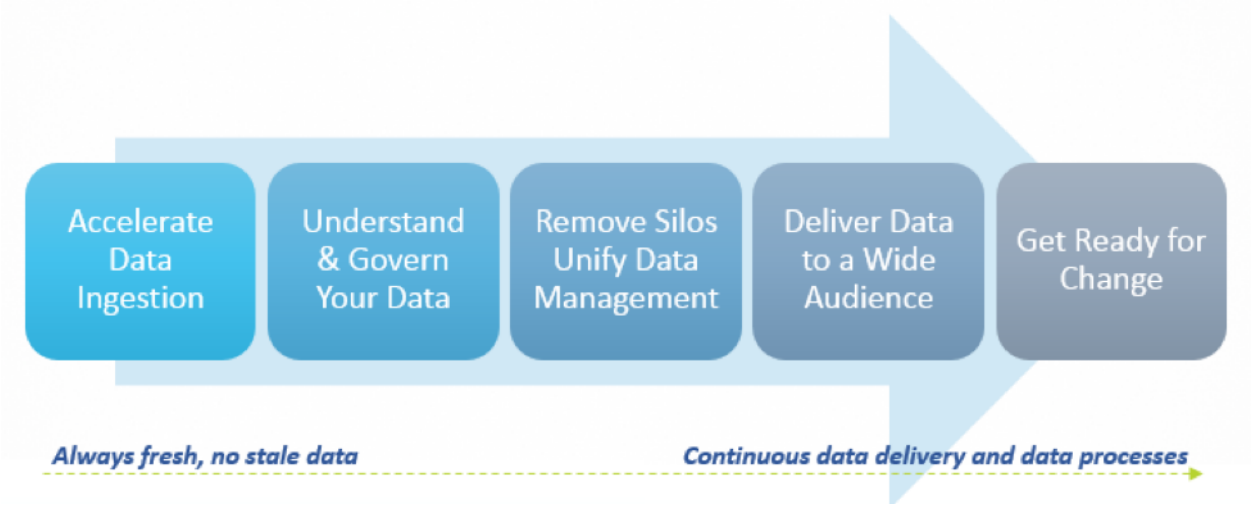
In the past, companies thought they'd gain full 360-degree visibility into their enterprise information with a data warehouse. However, the advent of big data has put these systems under distress, pushing them to capacity, and driving up the costs of storage. As a result, some companies have started moving some of their data (often times less

utilized data) off to a new set of systems like those run in Hadoop, NoSQL databases or the Cloud.

As a result of this migration, companies also came to realize that they can actually do more with Hadoop, NoSQL and Cloud vs. using enterprise data warehouses. Thus, they started adding new sources of data like sensor, mobile, social and big data to these systems, ultimately transforming their Hadoop, NoSQL and Cloud systems into data lakes.

According to Nick Huedecker at Gartner, "Data lakes are marketed as enterprise-wide data management platforms for analyzing disparate sources of data in its native format. The idea is simple: instead of placing data in a purpose-built data store, you move it into a data lake in its original format. This eliminates the upfront costs of data ingestion, like transformation. Once data is placed into the lake, it's available for analysis by everyone in the organization."

The data lake metaphor emerged because 'lakes' are a great concept to explain one of the basic principles of big data. That is, the need to collect **all** data and detect exceptions, trends and patterns using analytics and machine learning. This is because one of the basic principles of data science is the more data you get, the better your data model will ultimately be. With access to all the data, you can model using the entire set of data versus just a sample set, which reduces the number of false positives you might get.



The lack of data governance is preventing many organizations from fully opening up data lakes for all employees to use, because more often than not, data lakes contain sensitive data like social security numbers, date of birth, credit card numbers, etc. that need to be protected. Hence these organizations will not reap the full benefits and get their full return on data lake investment without having a thorough information governance strategy.

https://www.talend.com/blog/2016/12/21/your-resolution-list-for-2017-5-best-practices-for-unleashing-the-power-of-your/

Charting the Course Toward Value-Based Care With a Healthcare Data Lake

The large volumes of unstructured and semi-structured data that are currently siloed in EHRs, PACS, and lab systems will also need to be integrated to guide informed data-driven decisions. This enterprise-wide approach helps reveal actionable insights about an organization's performance indicators and impacts of patient care interventions to better manage risks and deliver affordable, higher quality care.

A data lake offers healthcare organizations a powerful data architecture with one, unified location for all healthcare data required for mining and analysis by clinical departments, business analysts, and data science teams. The end goal — reduce time to insights — is attained through the ability to analyze multiple variables to quickly identify trends, patterns, and correlations.

Complementing existing business intelligence and data warehouse investments, a data lake enables healthcare providers to execute analytics across disparate systems — running databases, data warehouses, and structured or unstructured data sets without impacting day-to-day operations or access to data.

https://www.healthitoutcomes.com/doc/charting-the-course-toward-value-based-care-with-a-healthcare-data-lake-0001

Diving in With a Healthcare Data Lake for Predictive Care

Not only is data coming into the healthcare system at a rapid rate, but it is also coming in many forms including structured, semi-structured, and unstructured formats - such as the EMR, patient images, lab reports, pathology, genomics, clinical notes, and social media activity. Instrumentation, sensor, and telemetry data. Yet, all of the information sources residing across the health system relevant to a particular patient care episode needs to be available to the right caregiver at the right time for a 360-degree view of the patient to provide a safe and appropriate diagnosis.

Moving toward predictive analytics creates future-focused insights-creating a new realm of data science for uncovering trends, patterns, relationships, correlations, and discoveries that impact integrated patient care. Data can also be consolidated from outside resources, including payers, genomic research centers, biobanks, and social media feeds.

Your organization can incorporate current business intelligence reporting, build internal data science capabilities, and then, deploy a data lake for healthcare to move forward predictive analytics and data mining.

Simply put, a data lake provides an IT environment that incorporates structured, semi-structured, and unstructured data from trusted external and internal sources, and ultimately improves effectiveness and quality of critical business and clinical practices. In addition, a data lake applies advanced analytics to produce actionable insights, enabling timely interventions to prevent adverse health events and ultimately, elevate overall population wellness.

https://www.linkedin.com/pulse/diving-healthcare-data-lake-predictive-care-eric-newman

Why Healthcare Organisations Should Take the Plunge With Data Lakes

Herein lies the solution in a data lake. Data lakes can benefit healthcare organisations by revealing actionable insights about an organisation's performance indicators and impacts on patient care interventions to better manage risks and deliver affordable, higher quality care. This type of technology draws all data into a single location, reducing or eliminating siloes across the healthcare enterprise, while collecting data from trusted external sources, such as research centres and public health databases.

The data lake makes it easy for healthcare professionals to mine and analyse all sources of data for insights into patient care. They can also analyse data to understand how they can make their organisations more efficient and in a way that won't adversely affect patient care. Analytics can provide insight to optimize staffing schedules and better manage IT resources, whilst helping to identify trends in services.

And with patient concerns over the privacy of information gathered, the value of actively managing this data also supports a number of other benefits including compliance with regulatory requirements, eliminating backup concerns and ensuring patient data is secure. As data enters the "lake," each pieces of information is identified with a range of security information that embeds security capability within the data itself. This type of approach could reduce barriers to information sharing between the industry and the public.

https://www.linkedin.com/pulse/why-healthcare-organisations-should-take-plunge-data-lakes-norman

Partners Data Lake Offers Healthcare Analytics as a Service

"We're putting public domain data in there, as well as institutional data," he said. "And then investigators can bring their own data, too. That's what's unique about the platform. It provides tools, storage, and technology, and the security wrapped around it, so investigators don't necessarily have to think about the infrastructure. But they can bring their ideas and begin to leverage those tools to develop their area of interest."

In the future, the platform will be able to ingest imaging data from radiology and pathology sources, as well, which will help to cultivate a rich and robust common data set that researchers draw upon for their work.

Data lakes are becoming a popular way to store massive volumes of healthcare data that may or may not have a clear use case at the moment. Unlike traditional relational databases, data lakes do not need to lock data into any particular format, category, or framework. Instead, a graph database or data lake can apply a standardized tag to each data element, and the tags can be combined and recombined in endless ways to produce nearly unlimited insights.

https://healthitanalytics.com/news/partners-data-lake-offers-healthcare-analytics-as-a-service

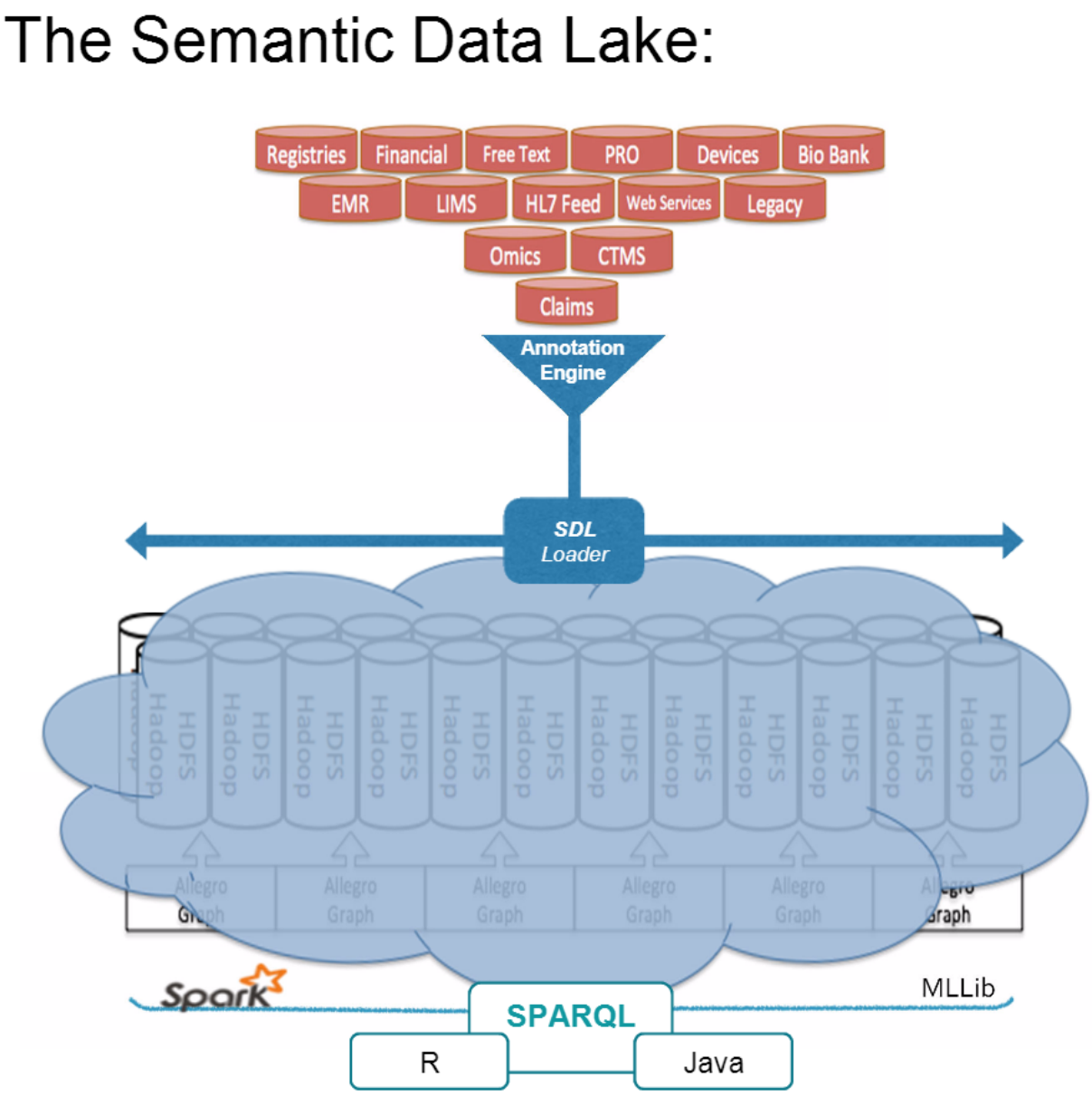
Medical Insight Set to Flow From Semantic Data Lakes

According to Franz CEO Jans Aasman, a semantic data lake employs a combination of technologies, including Hadoop, graph analytics, a semantic "triple store," the SPARQL query language, and Spark-based machine learning, to allow doctors to connect the dots between patient conditions and a world of knowledge contained in structured internal systems, as well as unstructured data sources outside of the organization.

Instead of hammering the data into a relational schema and storing it in a data mart, they want to pull all the pertinent data—including any "linked open data" such as drug interaction databases, genetic test results, or a demographics database sorted by ZIP code–into Hadoop and HDFS.

The semantic magic starts once the data is in Hadoop and HBase. Montefiore transforms all the data into semantic triples using a special ETL tool called a semantic annotation engine. Then the Franz AllegroGraph graph databases indexes all the triples and powers the SPARQL queries across the joined corpus, in addition to machine learning powered by Apache Spark ML or R.

As a result of this semantic approach, users can run queries that traverse different data sources, which was the stumbling block of traditional relational approaches.



https://www.datanami.com/2015/08/26/medical-insight-set-to-flow-from-semantic-data-lakes/