It’s a big deal: What 'big data' means to the health care system

Let’s Talk Informatics

June 30, 2022

Dr. Calvino Cheng, MD, PhD, FRCPC (HP)

Dalhousie University and Nova Scotia Health, Halifax, Nova Scotia

Informatics

Informatics utilizes health information and health care technology to enable patients to receive best treatment and best outcome possible.

This series is designed to enable participants to:

- Identify knowledge and skills healthcare providers need in order to use information now, and in the future.
- Prepare health care providers through an introduction to concepts and experiences in Informatics.
- Acquire knowledge to remain current by becoming familiar with new trends, terminology, studies, data and news.
- Collaborate with a network of colleagues to establishing connections with leaders who can provide advice on business issues, best-practice and knowledge sharing.

Continuing education credits

- **Digital Health Canada** - participants can claim 1CE hour for each presentation attended.
- **College of Family Physicians of Canada and Nova Scotia Chapter** - participants can earn one Mainpro+ credit by providing proof of content aimed at improving computer skills applied to learning and access to information.
- **Canadian College of Health Information Management -** approves 1 CPE credit per hour for this series for professional members of Canada's Health Information Management Association (CHIMA).

Conflict of Interest Declaration

- None to declare.

Session Specific Objectives

At the end of the presentation, the participant will be able to

- Appreciate the terminology of 'big data'
- Appreciate how 'big data' is stored
- Appreciate how 'big data' techniques could be applied to our health care system

Types of data in health care

- All of the various touch points as individuals interact with the system
  - Outside health system
    - Patient facing portals
    - Social media
  - Inside health system
    - Clinical services, laboratory, pharmacy, radiology, and other clinical information system data
    - Administrative (financial data)
    - Research (clinical trials)
    - Devices (wearables, Internet of Things, monitors)
- Storage of this data is typically done using a ‘relational database’

Storage of traditional datasets

- Relational Database (i.e. MySQL, SQLite, PostgresSQL) [RDBMS]
  - data is stored in tabular format, with schema defining tabular relationships, fields, and relationships; data stored in a column or an attribute

Product table

Product ID	Donor ID	ABO	Rh	Volume

Donor table

Donor ID	Donor name	Donor address	Gender	Birthdate

https://www.oracle.com/database/what-is-a-relational-database/

Queries from traditional datasets

- Structured Query Language (SQL) used to update, delete, create, retrieve, etc. data in relational database
  - E.g. find donors who have donated 100 apheresis platelets

Product table

Product ID	Donor ID	ABO	Rh	Product type

Donor table

Donor ID	Donor name	Donor address	Gender	Birthdate

https://www.oracle.com/database/what-is-a-relational-database/

Traditional enterprise architecture

- Transfusion 'chain'
  - Collection site (site ID, location, type of clinic – permanent vs. mobile, etc.)
  - Donor (donor ID, demographics, blood group, donation characteristics, pre-donation questionnaire answers, etc.)
  - Blood product (product ID, blood group, phenotyping, expiry, diluent, etc.)
  - Recipient (hospital ID, demographics, blood group, transfusion reactions, blood products, etc.)
  - Date- and time-stamps for all
- Other datasets relevant to the transfusion ecosystem
  - Twitter, Instagram, Facebook, and other social media platforms CBS interacts with (JavaScript Object Notation – JSON files)
  - Flat files from telephony and call center logs (i.e. chat files, chat logs)
  - XML files from subsidiary companies (i.e. equipment vendors and middleware)

Databases and datasources

→ ETL tools (Extract, Transform, Load) →

Data warehouse (non-customer facing; OLAP online analytical platform; can use analytics/visualisation tools on it; not public-facing)

Extract: Retrieves, verifies from sources
Transform: processes and organises into usable format
Load: moves transformed data into repository

https://www.informatica.com/resources/articles/what-is-etl.html

Problems with traditional enterprise architecture

- ETL tools are typically on a single machine
- Not real-time as typically ETL tools run at night or with ops jobs (can’t push out customized real-time offers)
- Transactions typically happen in data sources that a data warehouse cannot access
- Data warehouses are very costly and not very scalable

Problems with relational databases and larger complex datasets

- Scalability problem – with large datasets, join queries fail or have to wait
- Data must be divided or partitioned when get big; larger datasets queried slowly
- Data must be normalized, but to get data from different tables, a join query must occur
- Structured data (row/columns) can only be represented (no images, no audio)
- Price!

The solution?

Use 'big data' techniques

Origins of 'big data'

- Doug Laney (Gartner consultant, 2001)
  - 3 V attributes
- McKinsley Global Institute report (May 2011)

https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx
http://Searchdatamanagement.com

Definitions of 'big data'

- Characteristics (the 3 V's)
  - Volume (large amount of data, typically terabytes or more)
    - Social media, mobile phone and apps, emails, sensors, e-commerce, finance, weather, etc.
  - Variety (different data types)
    - audio, video, images, text, metadata
    - Structured (i.e. relational data, rows/column, tables), semi-structured (e.g. log files), unstructured (i.e. video, free text, pictures)
  - Velocity (data updated at fast pace, real-time data, data at rest)

Doug Laney, Meta Group Inc (Gartner), 2001

Definitions of 'big data'

- Characteristics (some more V's)
  - Veracity (accuracy of data, trustworthiness, applicability)
    - Garbage data = garbage results
  - Value (understanding the value of these big datasets)
    - Data is valueless until information is extracted from it
  - Variability (multiple meanings or formats data can have)
  - ….apparently, there are up to 10 V's.

- Datasets whose size is beyond the ability of traditional database software tools to capture, store, manage and analyze.
  - Intentionally subjective
  - Varies by sector and depends on software tools available

McKinsley Global Institute report (May 2011)
Doug Laney, Meta Group Inc (Gartner), 2001
Searchdatamanagement.com

Enablers of 'big data'

- Enabled by cheap storage and data generation, everywhere
  - Cheap hard drives can store the planet’s music
  - Mobile phones
  - Social media platforms
  - 235 terabytes of data collected by US Library of Congress (2011)
  - 15 of 17 sectors have more data stored per company than US Library of Congress
- Cheap computer processing, high speed networks, virtualization, cloud computing
- Sensors everywhere, Internet of Things, data created as a byproduct of other activities (i.e. “exhaust” data)
- Consumers and data generation when they communicate, browse, buy, share, search in this digitized world

Enablers of 'big data'

- New database platforms and architectures (i.e. Hadoop, MapReduce)
- Managed 'big data' platforms – cloud service providers such as AWS (Elastic MapReduce services, simple storage services, etc).
- Open source software (openstack, PostGresSQL)
- Funding (March 2012) – Obama announced $200M for 'big data' research.

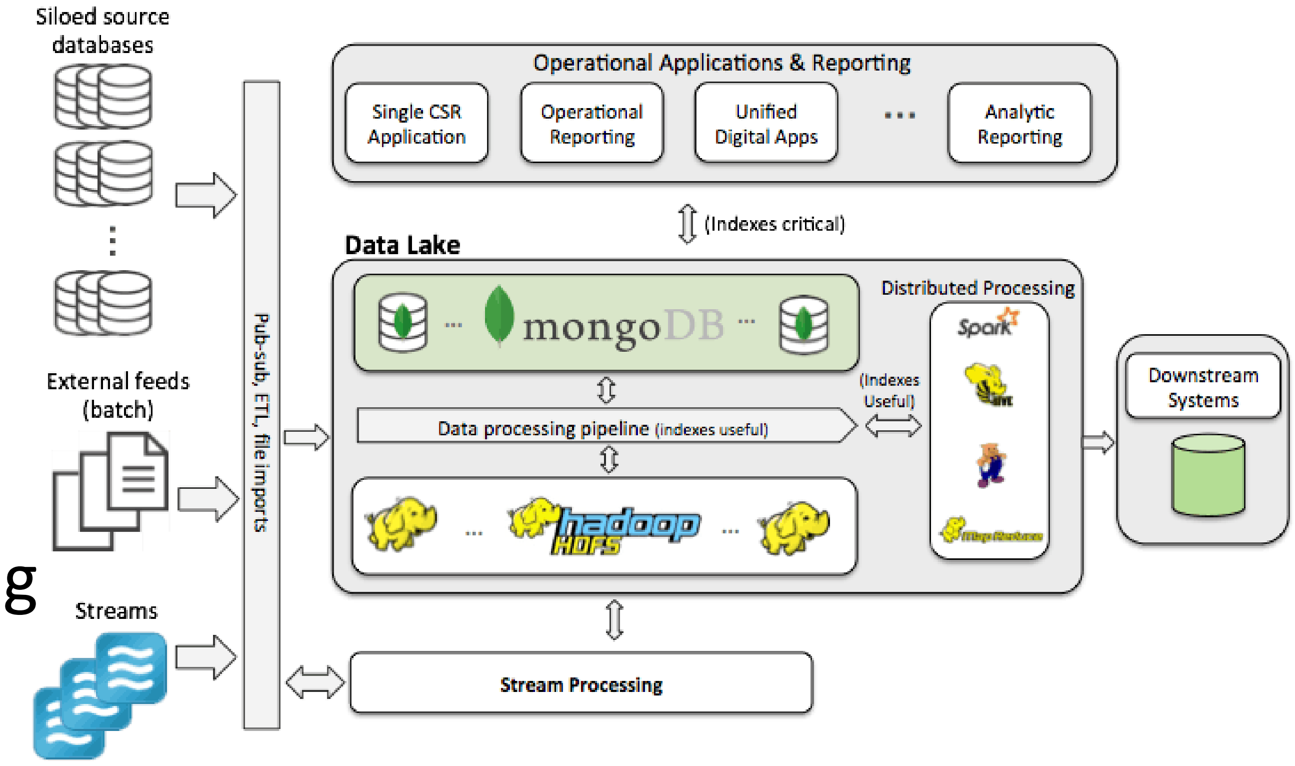
Enablers of 'big data'

- No SQL – Not Only SQL (non-SQL databases, use Python, Ruby, C, etc for retrieval); not limited to rows or columns; can store and retrieve unstructured data
- NewSQL – overcome MySQL scaling limitations; used in distributed processing
- Further database types: Key-Value Pair, document, columnar, graph, spatial, in-memory, cloud.
- Google file system (chunks of data stored, master server keeps map of locations and files)
- BigTable (distributed storage built on Google File System; not available outside of Google)
- NOTE: transactional management databases are typically on RDBMS

How is 'big data' stored?

Examples of 'big data' storage framework

- Commodity hardware
- Distributed storage
  - Hadoop Distributed File System
- Distributed processing
  - MapReduce
  - Apache Spark
- Resource scheduler/job scheduling
  - Yarn



Stream icon from: https://en.wikipedia.org/wiki/File:Activity_Streams_icon.png

https://hadoop.apache.org/
https://www.mongodb.com/big-data-explained/architecture

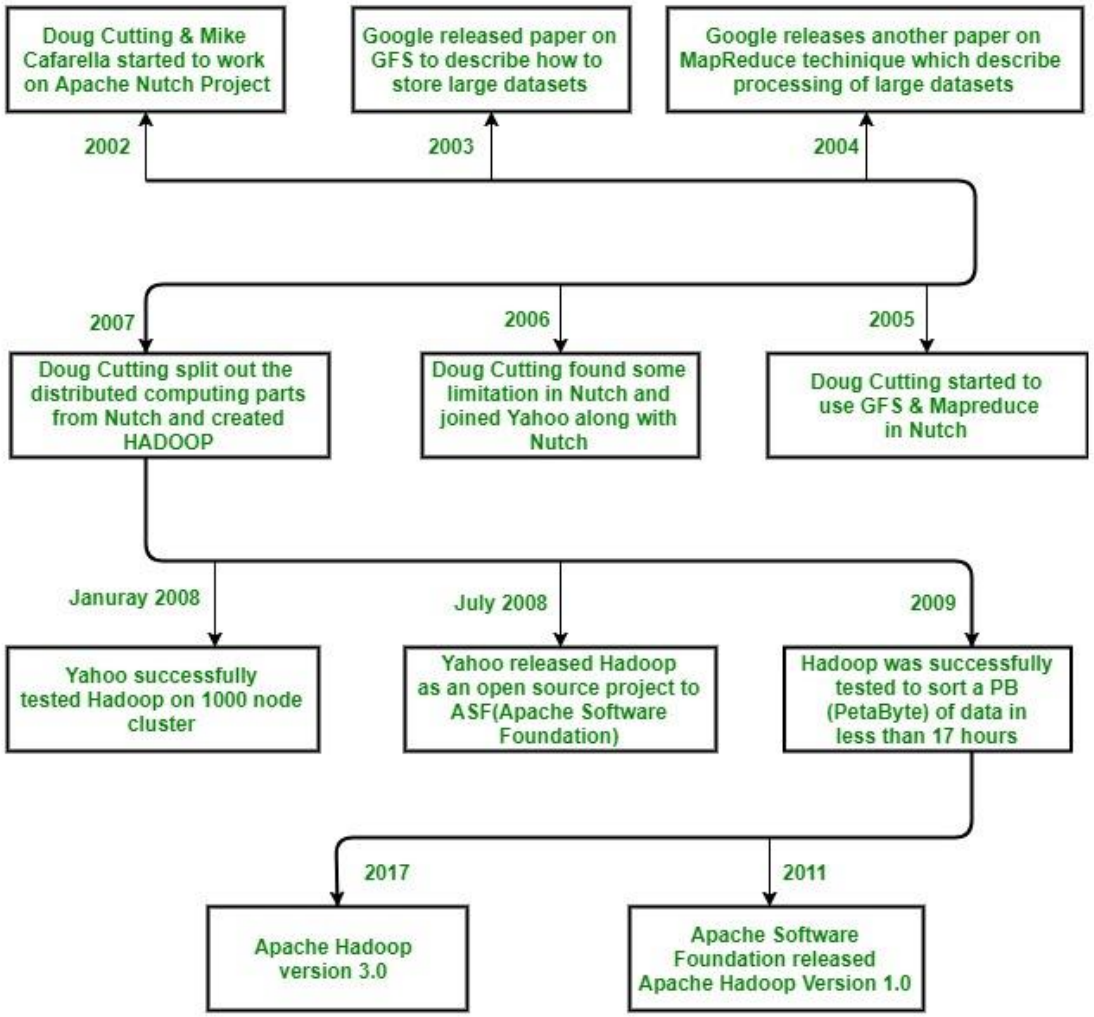
Advantages of Hadoop (Apache)

- Cheap – commodity hardware, open source
- Fault tolerance – HDFS stores 3 copies of data, multiple name nodes
- Scalability – scale up or down by adding or subtracting cluster hardware
- Distributed processing - faster
- Data locality – sending a local query over to a remote dataset and fetching the result, rather than bringing data to the local machine.
  - Vs. traditional data and query processing
    - Data stored on local machine, query processed on local machine
    - Data stored on remote servers, query processed on local machine

Hadoop (Apache)

- Designed by Doug Cutting and Michael Cafarella in 2005
  - Inspired by Google File System (GFS)
- 3 components of Hadoop
  - Storage of data = HDFS (Hadoop Distributed File System)
    - Stores data across multiple machines in the cluster using commodity hardware (data security, fault tolerance)
  - Processing of data = MapReduce
    - When query sent to process data, Hadoop knows where it is stored (i.e. Mapping)
    - Query sent to those machines and processed
    - Recombination of the queries sent back to user (i.e. Reduce)
  - Resource manager = YARN (Yet Another Resource Negotiator)
    - Job scheduler, resource manager (First Come First Serve, Capacity Scheduler, Fair Share Scheduler)

Evolution of Hadoop



https://www.geeksforgeeks.org/hadoop-history-or-evolution/?ref=rp

Evolution of Hadoop

- Hadoop 1 = Original Hadoop (Hadoop Common utilities; HDFS; MapReduce)
- Hadoop 2 = added YARN
- Hadoop 3 = most recent version (multiple namenodes, and other upgrades)

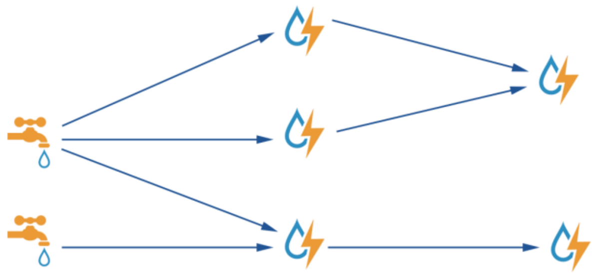
https://hadoop.apache.org/

Evolution past Hadoop: the new ecosystem

- Hadoop has matured, but it is still not the best solution
- Batch processing falling out of fashion – users want real-time processing
- Hadoop is still ‘complex’ and easier debugging is required

The new ecosystem

- Apache Spark
  - Initially created to batch process, attached to Hadoop; no need for Hadoop now
  - Supports stream processing (real time data processing, helps with AI applications) using in-memory processing, rather than disk-based (100x faster)
- Apache Storm
  - The equivalent for Hadoop, but for real-time data streams
  - Specialized for complex event processing (CEP)
  - Data analyzed in continuous stream vs. Hadoop (data must enter file system to get processed)



https://storm.apache.org/
https://spark.apache.org/

The new ecosystem

- Ceph
  - No single point of failure, completely distributed platform
  - Reduces administration costs (i.e. related to fixing errors on server clusters)
  - Ceph storage system scales better than HDFS for convoluted directory structures
- Hydra
  - Supports streaming and batch processing
- Google BigQuery
  - Fully managed, runs on Google’s hardware
  - Built-in data mining algorithms, runs complex queries, backwards compatible with MapReduce
  - Fast(hours in Hadoop, minutes in BigQuery); structured nature (more data control)

https://www.hdfstutorial.com/

How to generate value from 'big data'

'big data' Analytics

What is 'big data' analytics?

- The process of deriving value from large datasets by gathering, organizing, analyzing large datasets to discover patterns, correlations, and meaningful insights
- Done to improve business processes (increase efficiency, boost profits, increase customer satisfaction)
- Usually done by 'big data' analysts, data scientists, statisticians, etc.

https://www.educba.com/what-is-big-data-analytics/

Uses of 'big data' Analytics

- Risk management
- Product development and innovation
- Quicker and better decision making
- Improving customer experience

https://www.simplilearn.com/what-is-big-data-analytics-article

Lifecycle phases of 'big data' analytics

- Business case creation (defines goal for analysis)
- Identification of data sources
- Filtering (data cleaning to remove corrupt data)
- Extraction (data transformation to make data compatible with tool)
- Aggregation (same fields across datasets integrated)
- Analysis (use machine learning, analytical tools, statistical tools)
- Visualization
- Analysis

https://www.simplilearn.com/what-is-big-data-analytics-article

Types of 'big data' analytics

- Descriptive
  - Summarizing data and making it human readable (i.e. profit reports)
- Diagnostic
  - Data mining, drilling down, discovering the reason for a problem
- Predictive
  - Use of current and historical data to extrapolate the future, such as trends
  - Data mining, artificial intelligence/machine learning
- Prescriptive (<u>pre</u>dictive + des<u>criptive</u>)
  - Proposes solution to a problem
  - Uses artificial intelligence/machine learning
  - Uses predictive and descriptive analytics

https://www.simplilearn.com/what-is-big-data-analytics-article

What does this mean for medicine?

Transfusion medicine example:

- Transfusion ‘chain’
  - Collection site (site ID, location, type of clinic – permanent vs. mobile, etc.)
  - Donor (donor ID, demographics, blood group, donation characteristics, pre-donation questionnaire answers, etc.)
  - Blood product (product ID, blood group, phenotyping, expiry, diluent, etc.)
  - Recipient (hospital ID, demographics, blood group, transfusion reactions, blood products, etc.)
  - Date- and time-stamps for all
  - Molecular data (red cell, platelets, organs and tissues)
  - Phenotyping (human platelet antigen, HLA – see above)
  - Cost data
- Other datasets relevant to the transfusion ecosystem
  - Twitter, Instagram, Facebook, and other social media platforms CBS interacts with (JavaScript Object Notation – JSON files)
  - Flat files from telephony and call center logs (i.e. chat files, chat logs)
  - XML files from subsidiary companies (i.e. equipment vendors and middleware)

How 'big data' can help transfusion

- Use your imagination
  - Real-time donor selection based on donation frequency, GPS data, inventory data
  - Donor selection and retention
  - Realtime feedback for client and donor satisfaction
  - Integration of hospital transfusion data
  - Forecasting or predicting wastage at supplier or hospital level
  - Predicting units at risk for recall

How can ‘big data’ thinking help our health care organizations locally?

- Use your imagination
  - Could we rethink the way we do analytics?
  - Could we deploy cheaper consumer-grade analytics infrastructure and re-think how we house offline data?
  - Could we use real-time data and machine learning to change the way we practice?
  - Could all transactions destined for the interface engine be duplicated and stored on big data architecture for real-time use?
- We need infrastructure that will boost our research and analytics capability while being budget-friendly

Summary

- ‘Big data’ techniques are required for large amounts of data that cannot be stored, processed, or analyzed using conventional means
- Hadoop is an example of a 'big data' ecosystem
- ‘Big data’ analytics are required to derive value from large datasets
- As a health care institution, we need to think differently about how we can handle data in the future

Questions and discussion

calvino.cheng@nshealth.ca