

DEALING WITH DATA

SPRING 2019: INFO-GB.2346.30

PROFESSOR GUTHRIE COLLIN

TEACHING FELLOW AJINKYA WALIMBE



NYU | STERN


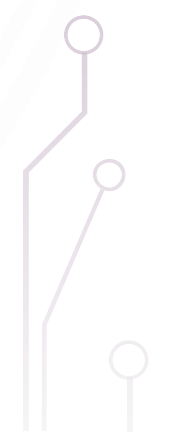


CLASS 11-B: BASIC BIG DATA CONCEPTS

MAY 2, 2019

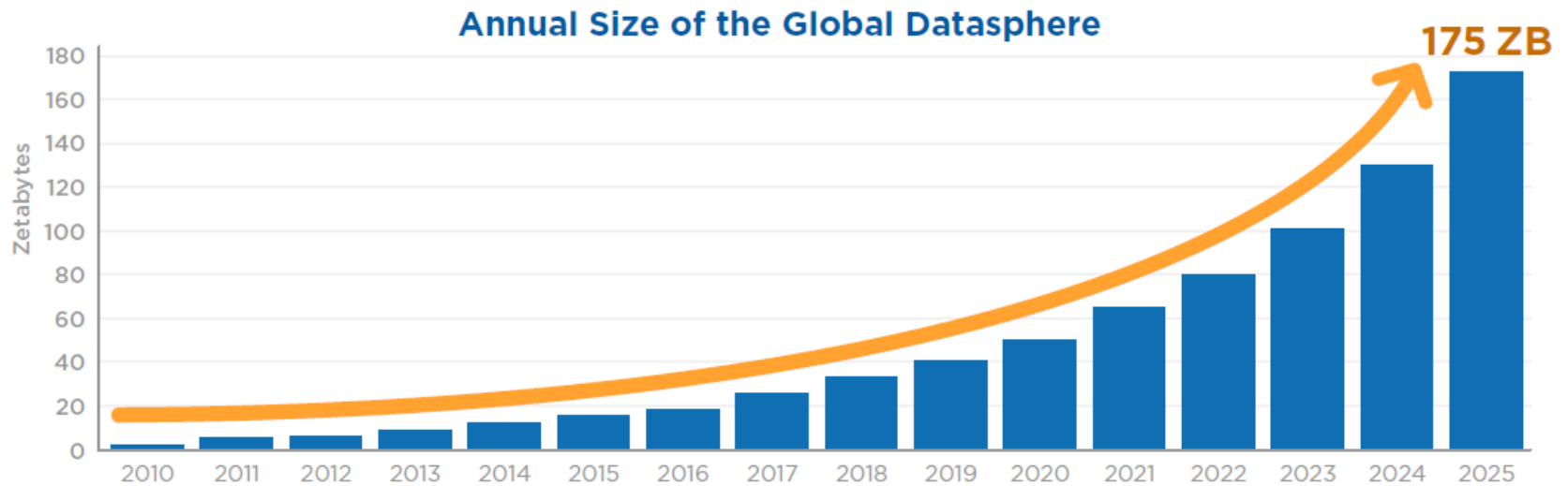


BIG DATA BASICS

1. Defining Big Data
 2. Hadoop and MapReduce
 3. Streaming Data
 4. NoSQL
- 
- 

DEFINING BIG DATA: INCREDIBLE DATA GROWTH

Figure 1 - Annual Size of the Global Datasphere



Source: IDC DataAge 2025 whitepaper

NOTE: 1 Zettabyte = 1 billion Terabytes

DEFINING BIG DATA: HISTORY

- 2001: Meta (now Gartner) first noted the increasing Volume, Velocity and Variety of data emerging as a consequence of digitization of our world.
 - This report didn't coin "big data", but the "3 Vs" framework became popular.
 - They have been joined by 4th and 5th Vs: Veracity (accuracy) and Value
- 2013: Univ of St. Andrews, UK researchers J.S. Ward and A. Barker attempt to unify a "big data" definition in an arXiv paper (109 citations since it was made public)

DEFINING BIG DATA: COMPETING VERSIONS CITED BY WARD & BARKER

- **Oracle:** *Big data is the derivation of value from traditional relational database-driven business decision making, augmented with new sources of unstructured data.*
- **Intel:** *Big data opportunities emerge in organizations generating a median of 300 terabytes of data a week. The most common forms of data analyzed in this way are business transactions stored in relational databases, followed by documents, e-mail, sensor data, blogs, and social media.*
- **Microsoft:** *Big data is the term increasingly used to describe the process of applying serious computing power—the latest in machine learning and artificial intelligence—to seriously massive and often highly complex sets of information.*
- **National Institute of Standards and Technology:** *big data is data which “exceed(s) the capacity or capability of current or conventional methods and systems.”*


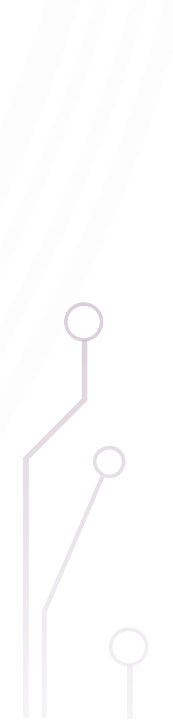
DEFINING BIG DATA: WARD & BARKER -VS- GARTNER

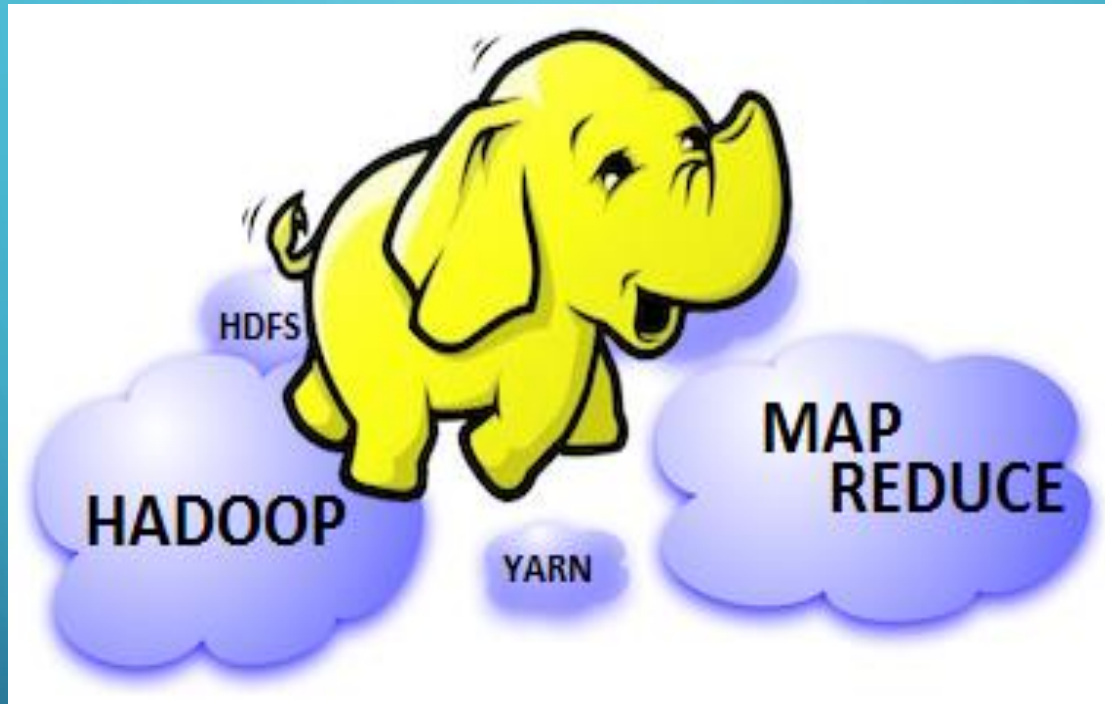
WARD & BARKER (2013): *Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.*

GARTNER (2019): *Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.*



BIG DATA: BUSINESS IMPLICATIONS

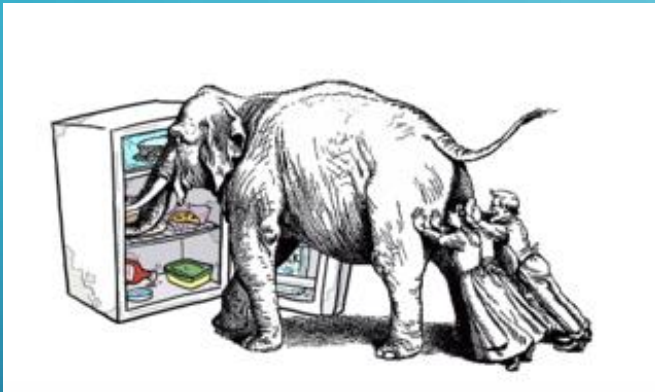
1. Requires significant tech capabilities + ongoing investment to handle Volume, Velocity & Variety
 2. Requires scientific rigor to enforce high Veracity
 3. Requires business vision and business+tech understanding to realize high Value opportunities
- 
- 



Credit for this section:
NYU's High Performance Computing Lab

BIG DATA CHALLENGES

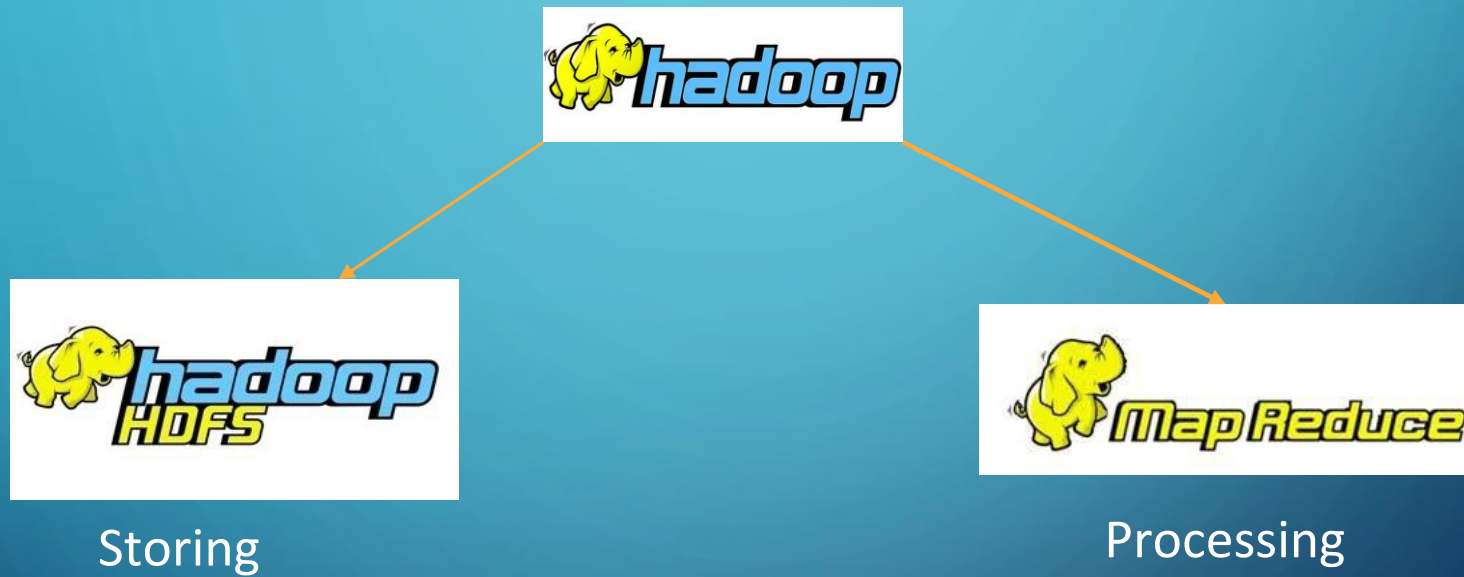
Storing Big Data

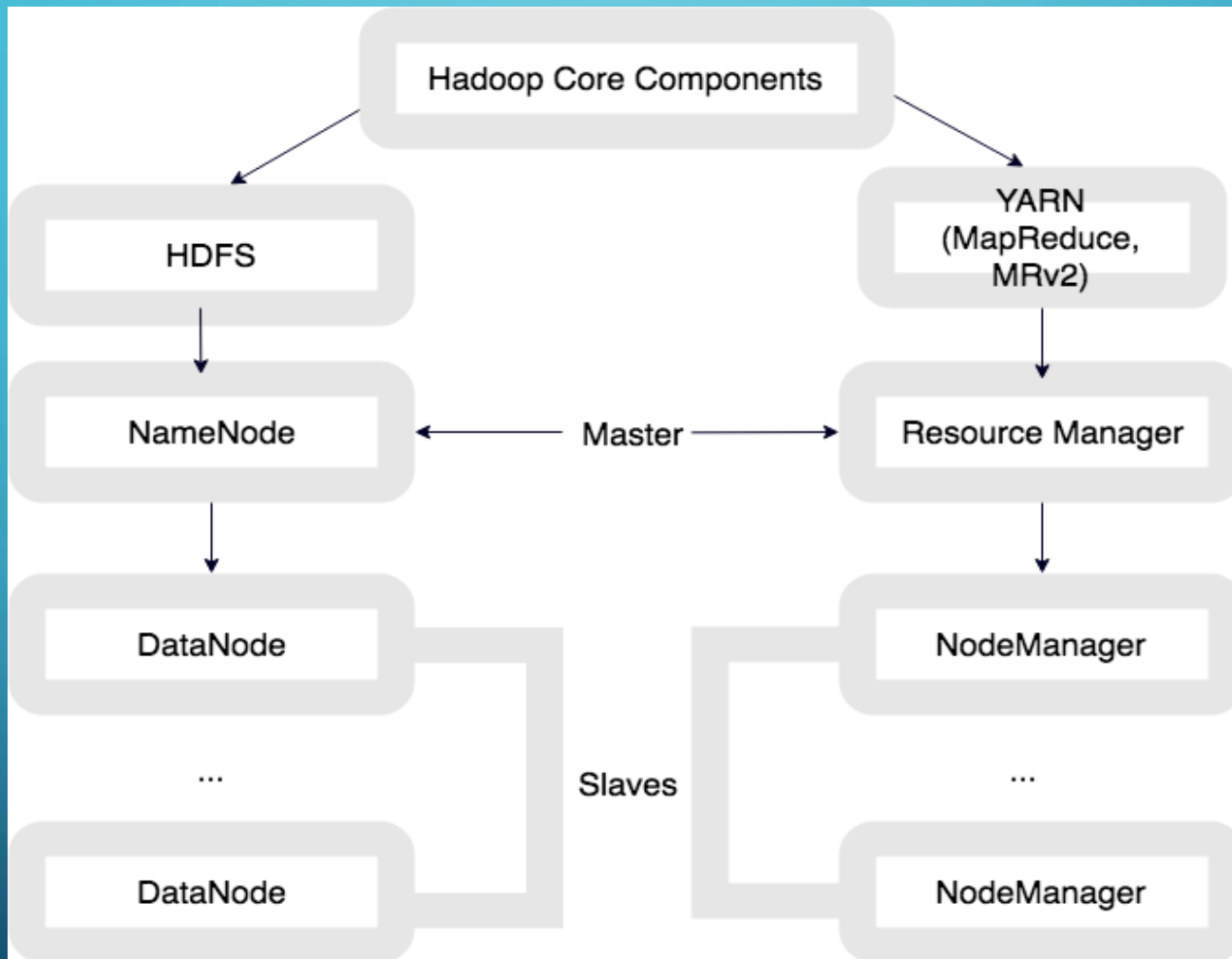


Processing Big Data



HADOOP MAIN COMPONENTS





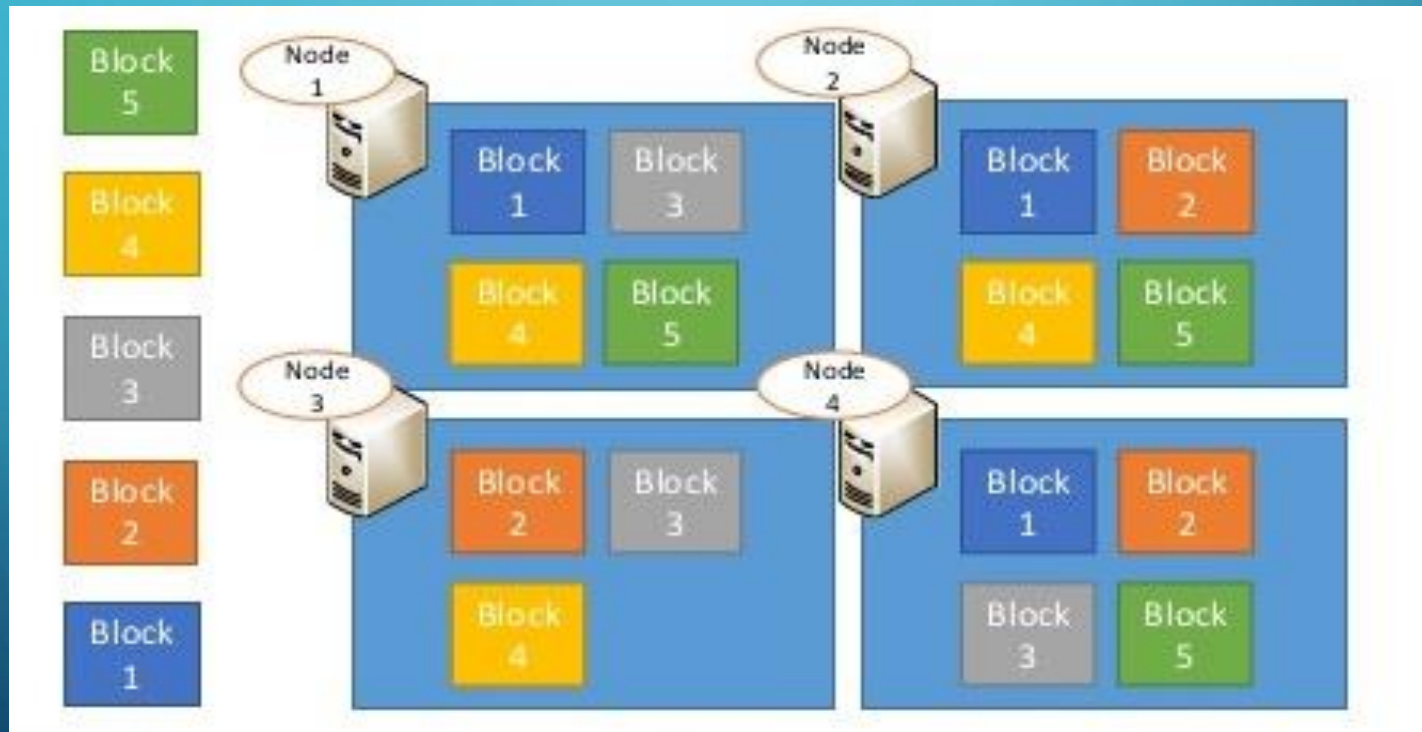
HDFS BLOCKS

- Each file is stored on HDFS in blocks
- Default Block Size in HDFS is 128 MB



HDFS REPLICA

- Default replica factor is 3
- Block placement is **rack-aware**

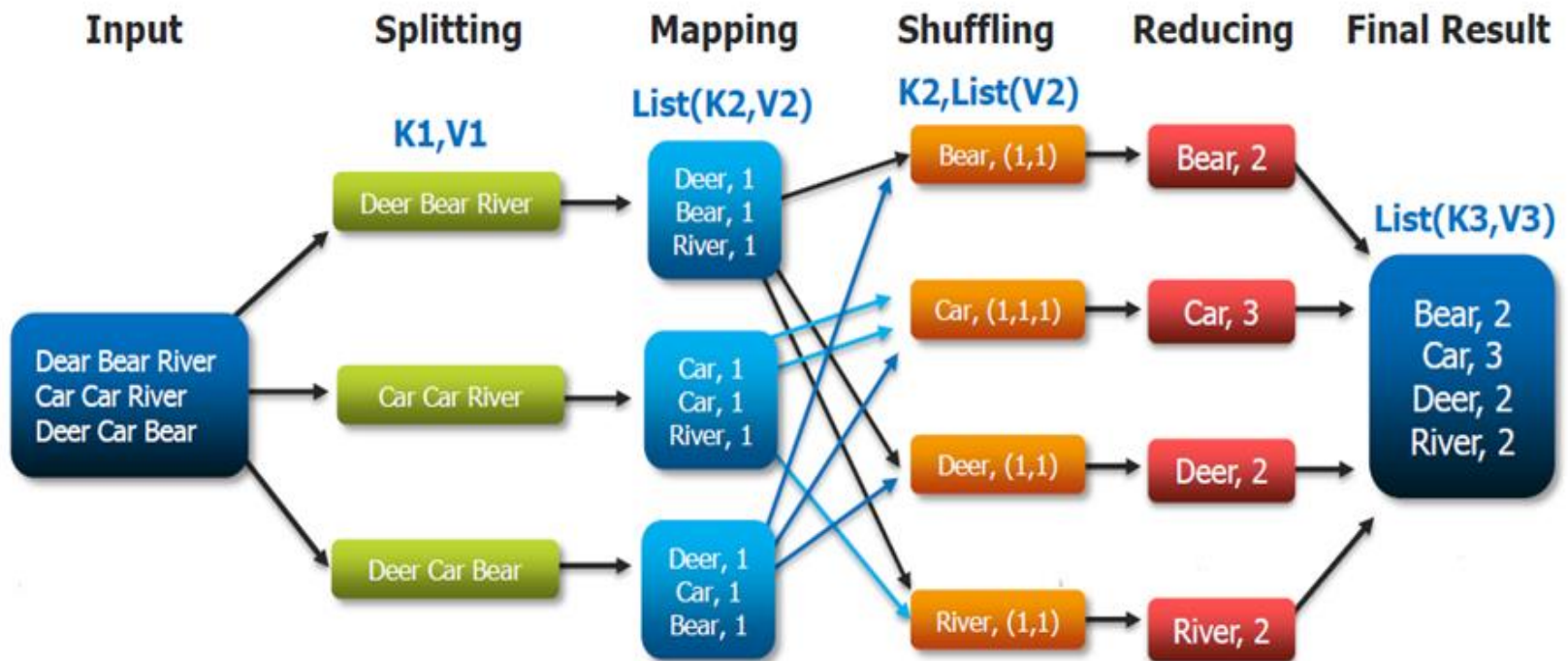


MAPREDUCE

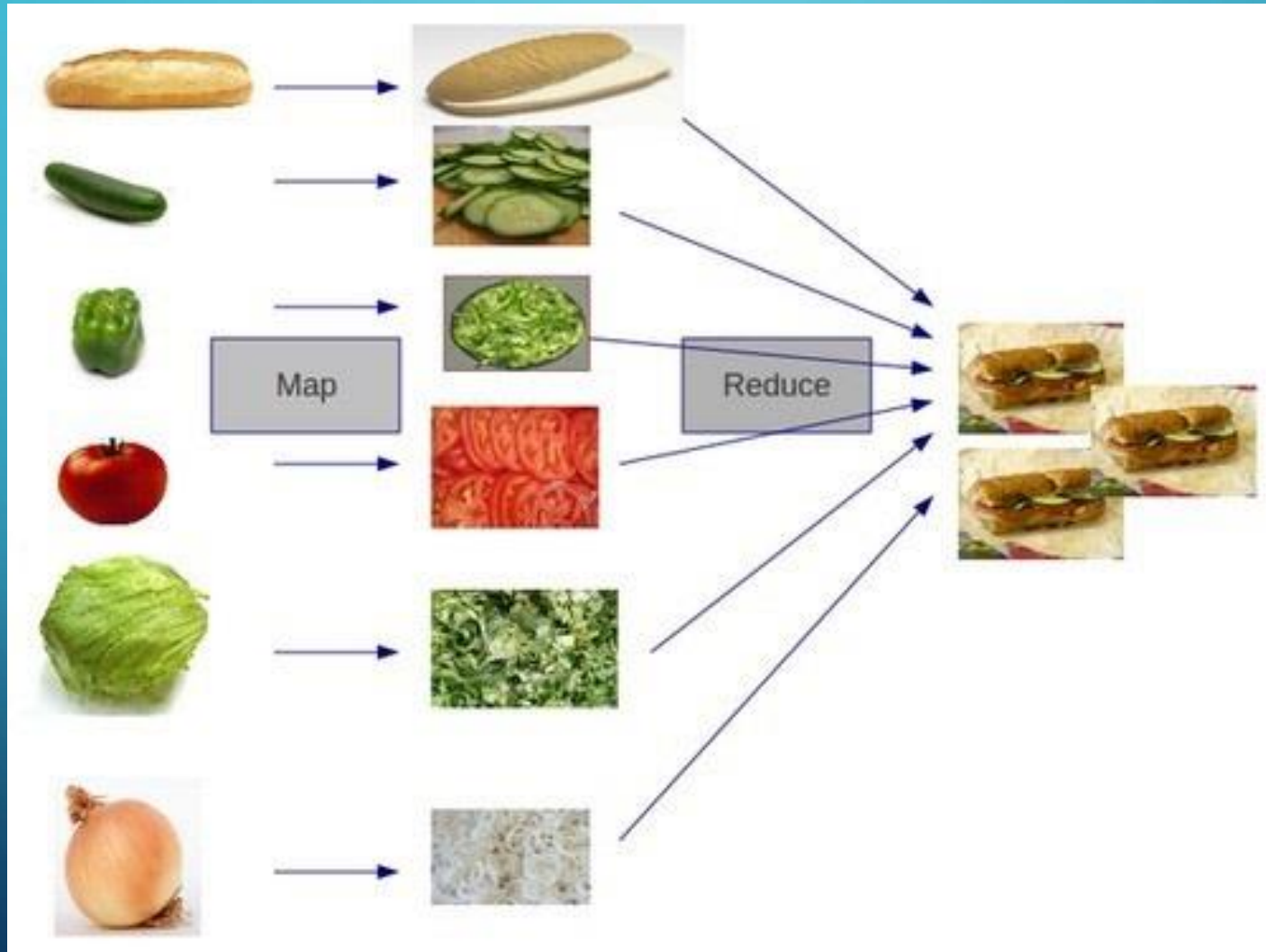
- MapReduce is a framework for processing parallelizable problems across large datasets using a large number of compute nodes
- It organizes data into key-value pairs (KVP) for parallel processing
- 3 steps of MapReduce:
 - Map step
 - Shuffle step
 - Reduce step

WORD COUNT EXAMPLE

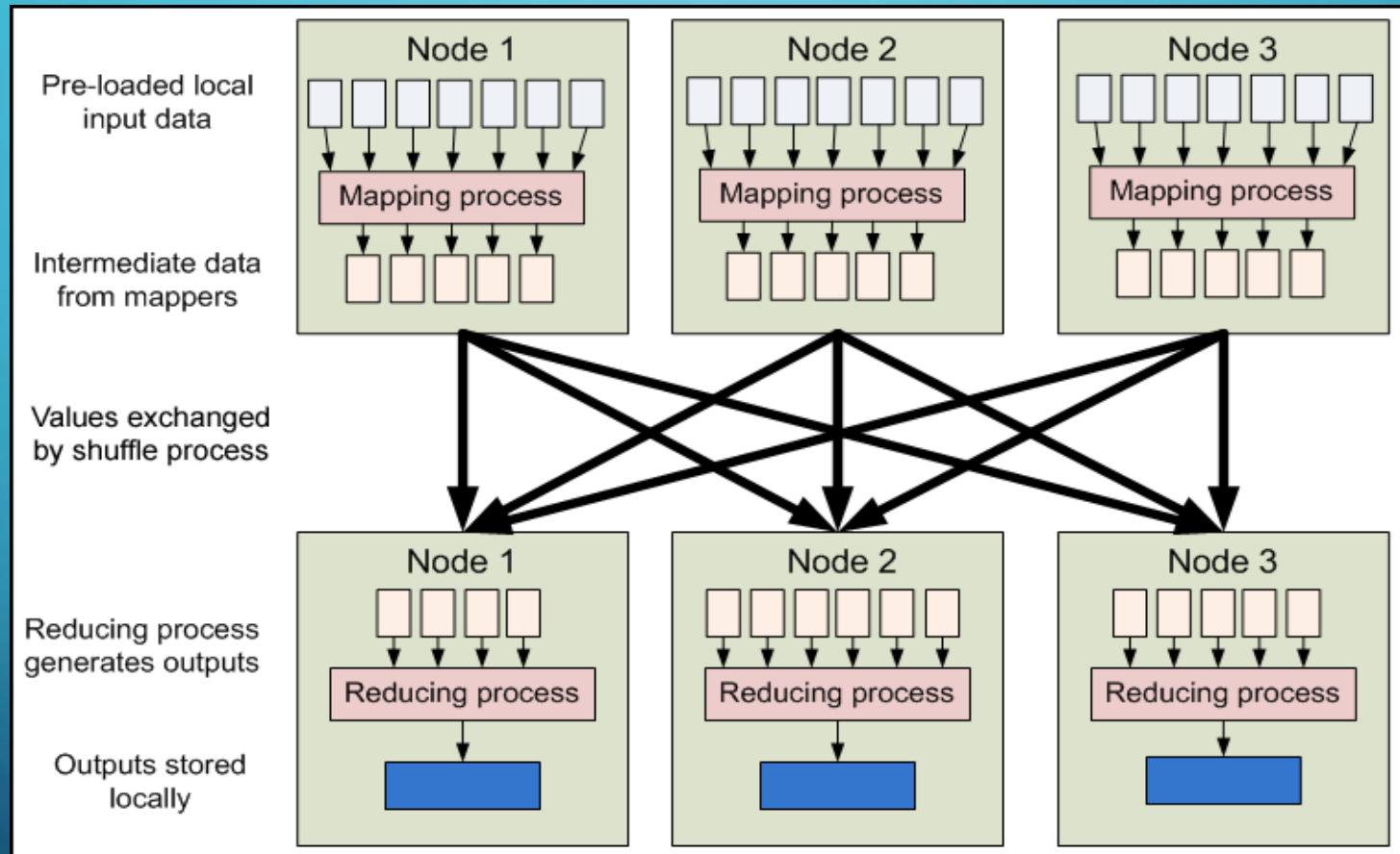
The Overall MapReduce Word Count Process



SANDWICH EXAMPLE



DATA FLOW

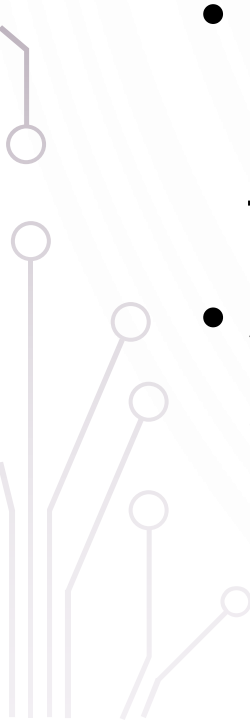
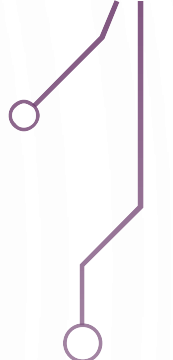
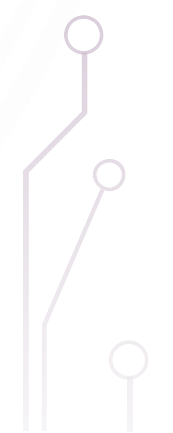


MAP REDUCE: BUSINESS IMPLICATIONS

1. Reduce dependency on specialized, expensive tech partners
2. Business decisions on required data processing speed are near directly correlated with cost:
X% faster data = approximately X% greater costs, and vice versa.



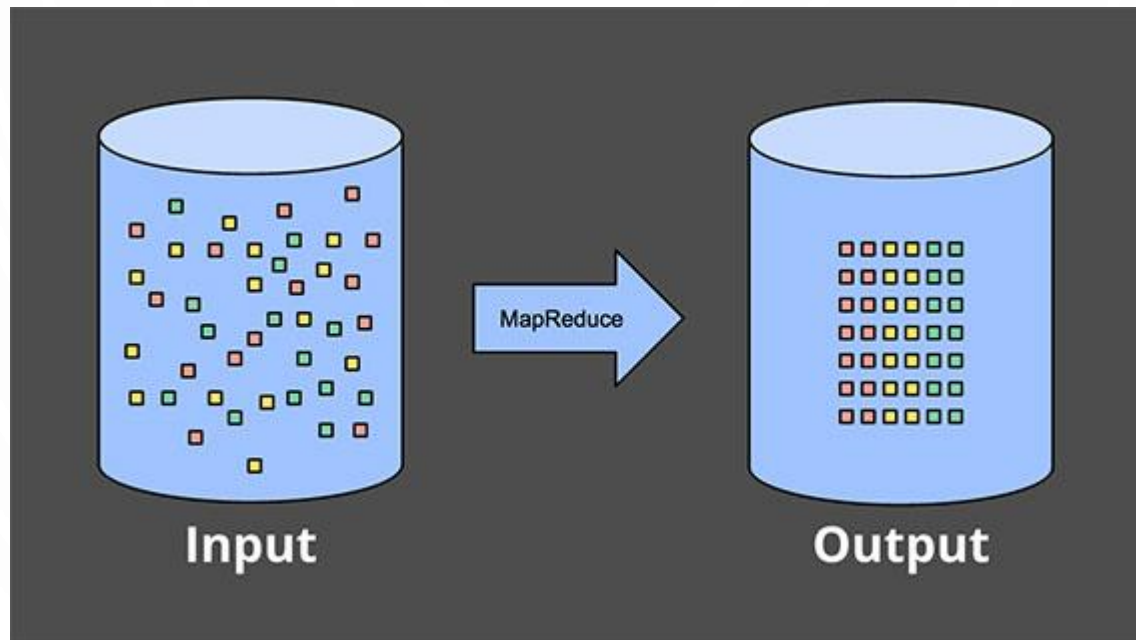
STREAMING DATA

- Data concept focused on working with unbounded/infinite data, e.g. ecommerce orders, tweets
 - Traditional database systems are designed to handle bounded data that is processed in batches, e.g. store all today's stock trading activity overnight
 - Streaming systems are designed to process unbounded data in real-time, sometimes in “micro-batches”
- 
- 
- 

STREAMING DATA: FAST! BUT, ACCURATE?

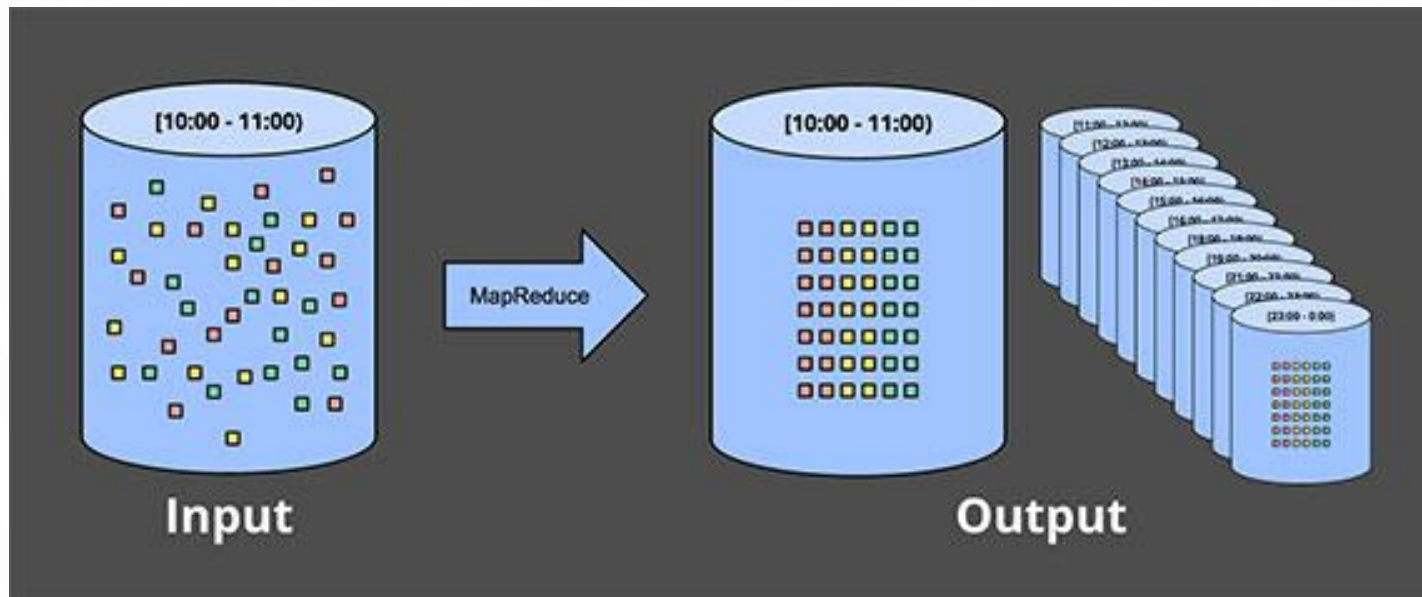
- Streaming data processing is ALWAYS FASTER than batch processing
 - Data processed immediately after the event versus “next day”
- Streaming data processing CAN BE AS ACCURATE as batch processing, but it depends on:
 - Are all the accuracy-ensuring processes capable of stream processing?
 - Online vs offline bot/fraud detection
 - Are there any business processes that can happen outside of the stream?
 - User cancelling an eCommerce order days after the purchase event
 - Credit card declined minutes/hours after the purchase event

BATCH DATA PROCESSING: ASSUMES FINITE DATA



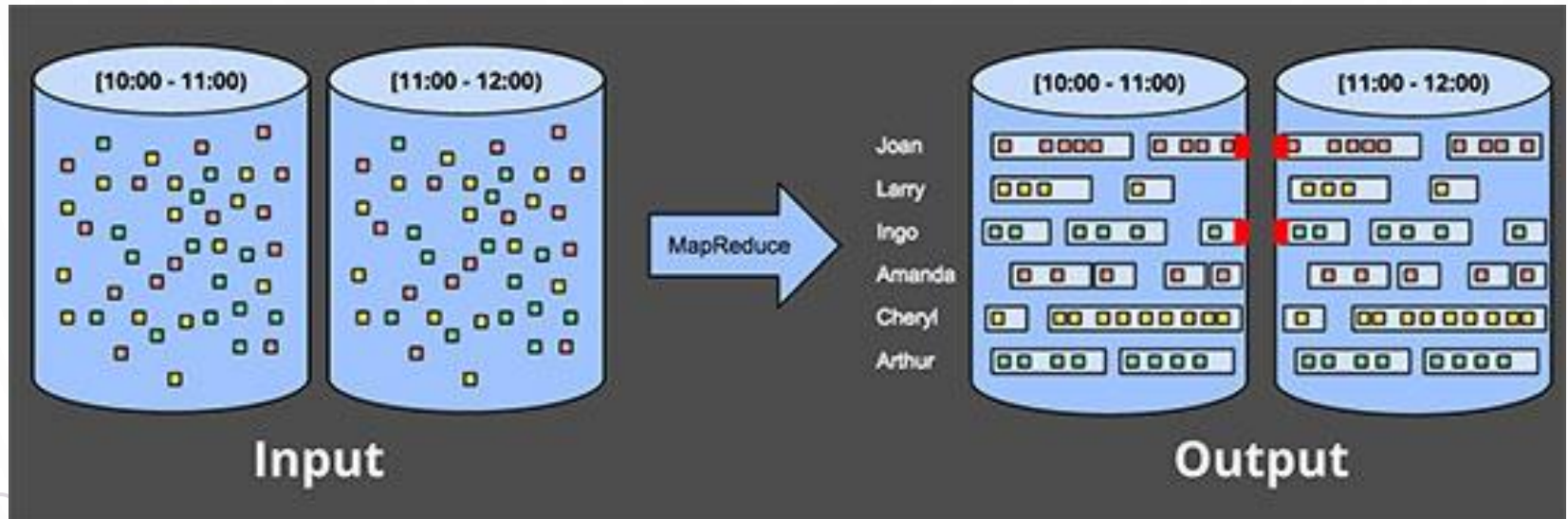
Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

BATCH DATA PROCESSING: WINDOWING INFINITE DATA INTO FINITE CHUNKS



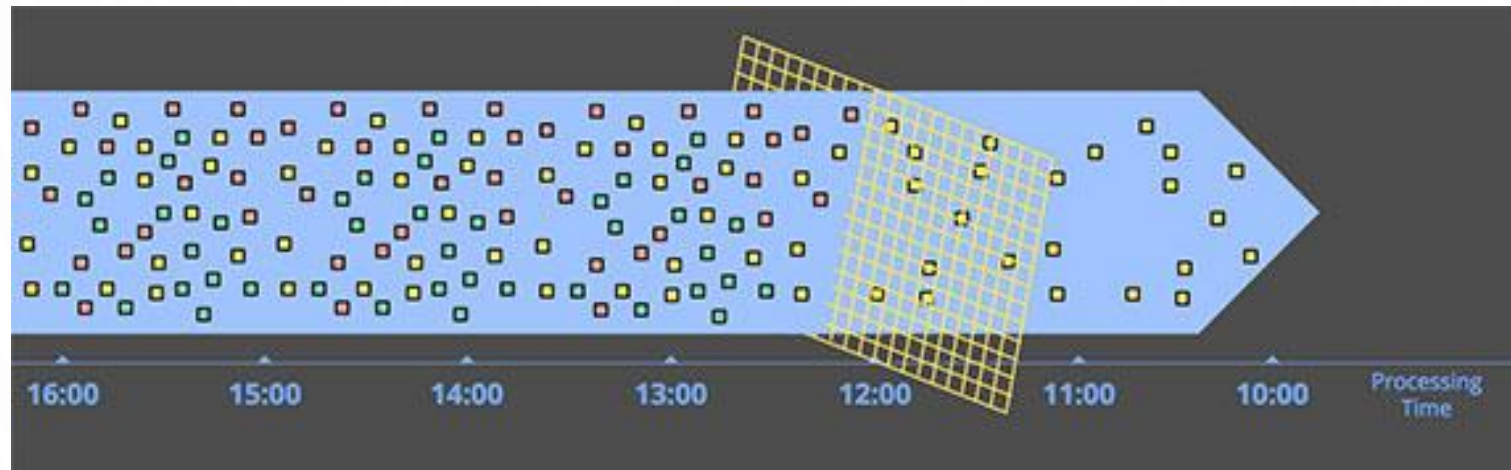
Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

BATCH DATA PROCESSING: WINDOWING INFINITE DATA INTO FINITE CHUNKS



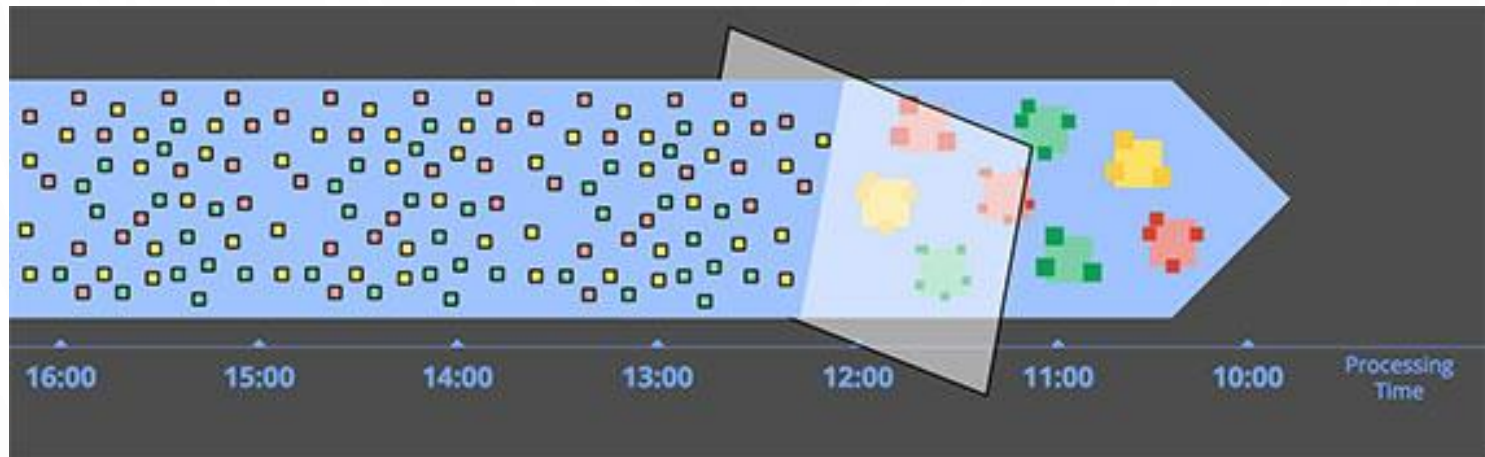
Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

STREAMING DATA PROCESSING: FILTERING THE DATA YOU WANT



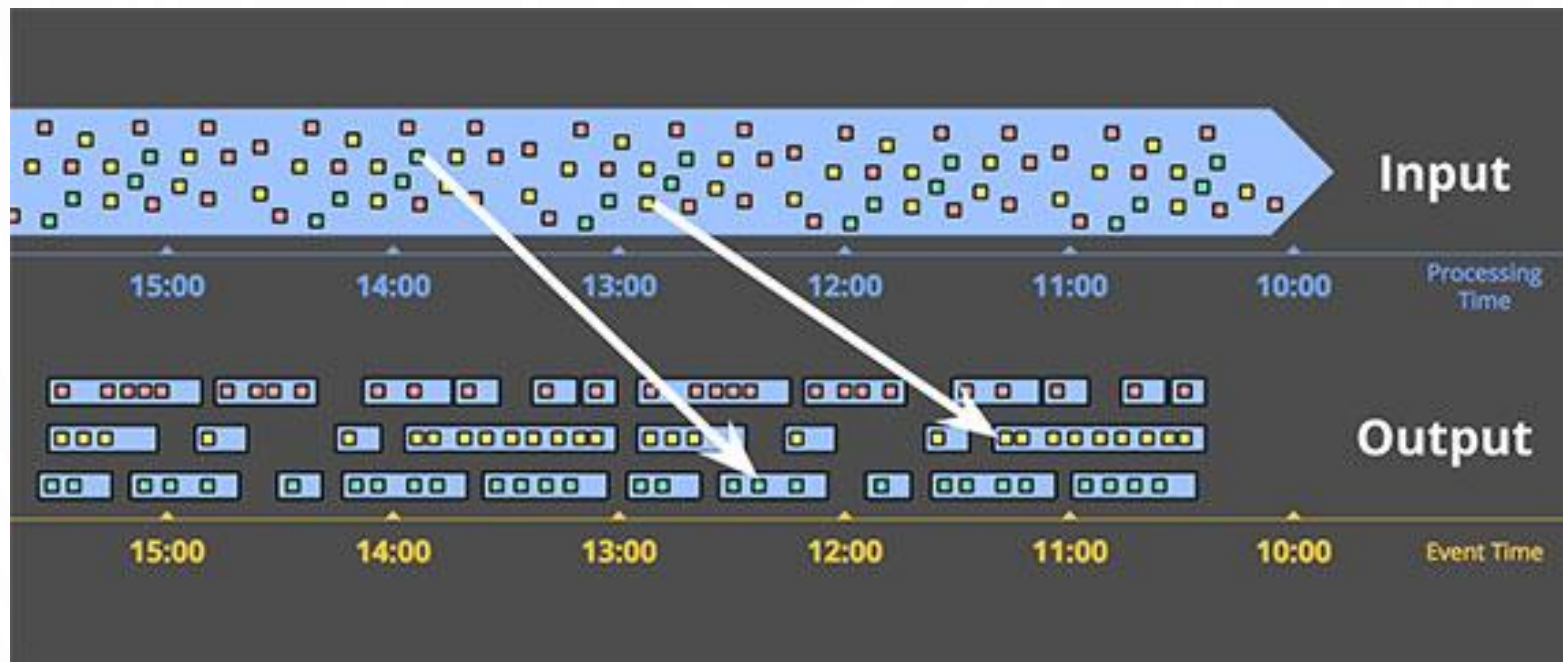
Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

STREAMING DATA PROCESSING: MACHINE LEARNING DATA ESTIMATES



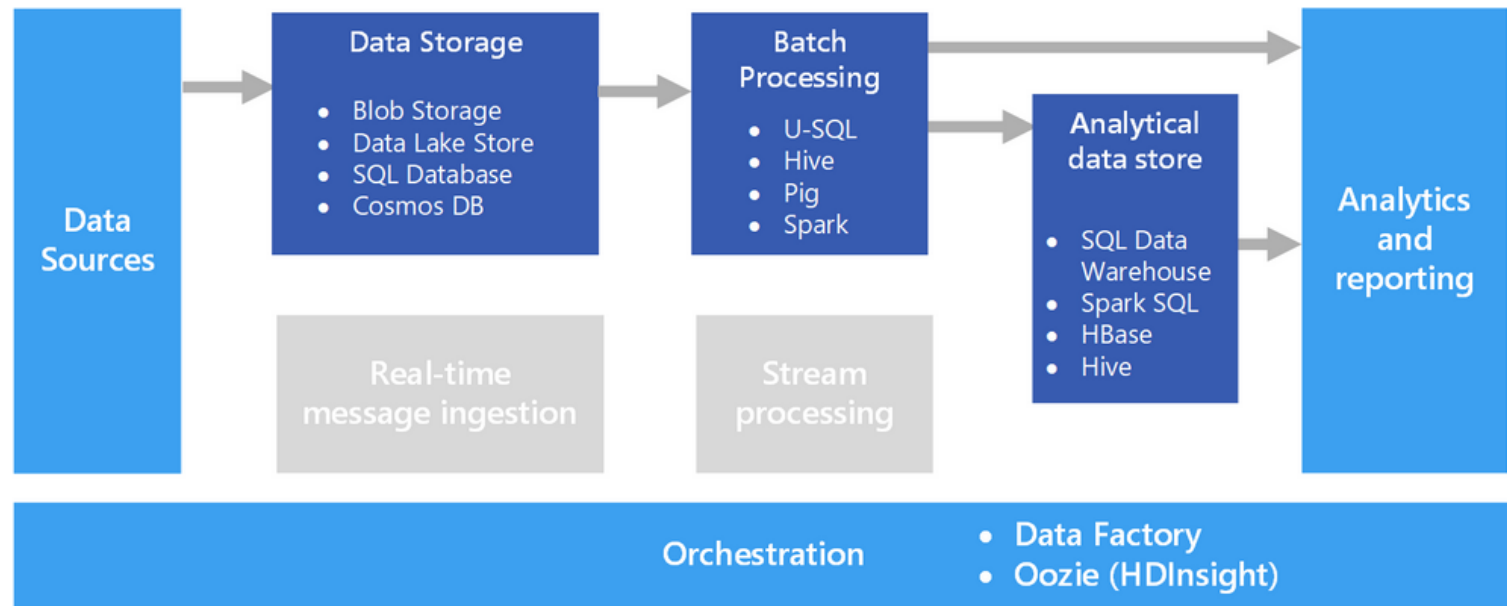
Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

STREAMING DATA PROCESSING: SLICING DATA BY USER OVER TIME



Credit: Tyler Akidau from <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101>

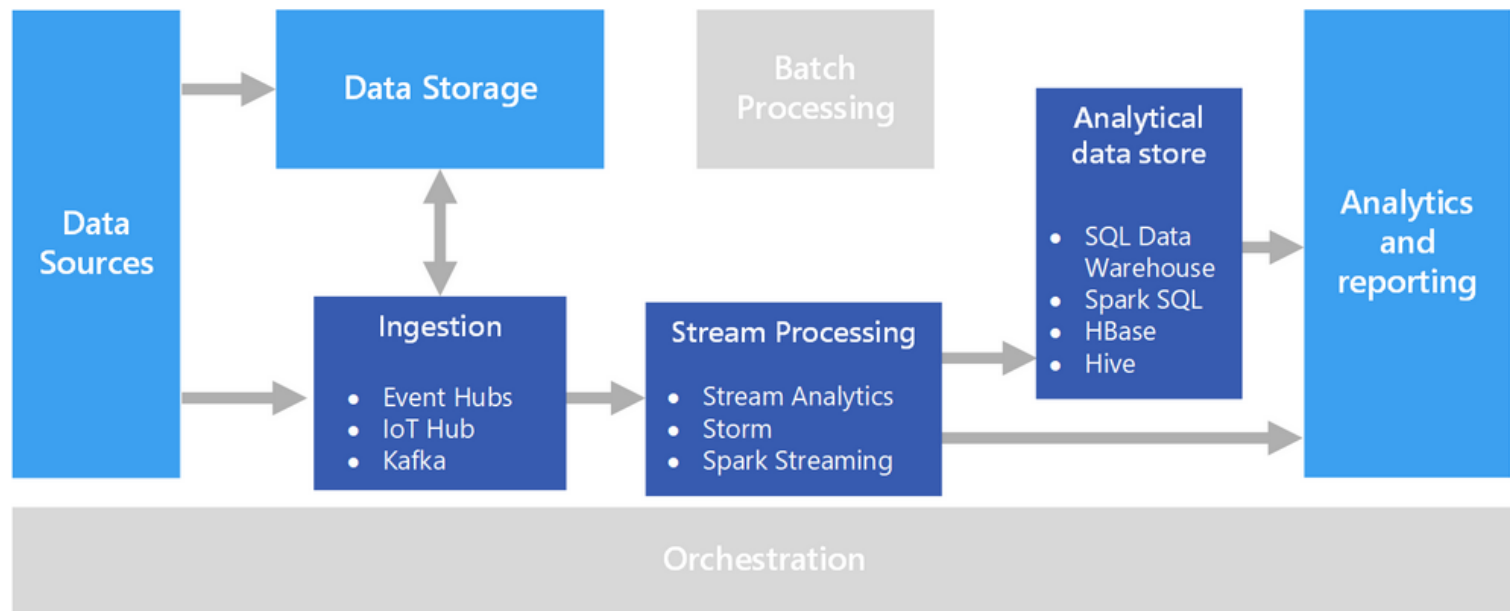
MICROSOFT AZURE: BATCH PROCESSING REFERENCE ARCHITECTURE



CREDIT: Microsoft Azure's Big Data guide ➔

<https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/batch-processing>


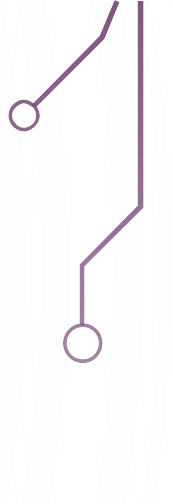
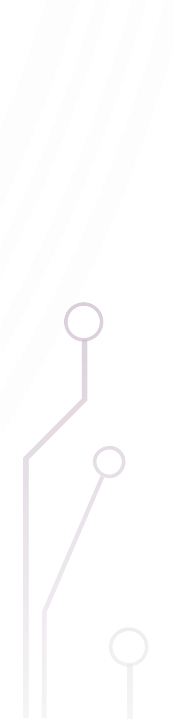
MICROSOFT AZURE: STREAM PROCESSING REFERENCE ARCHITECTURE



CREDIT: Microsoft Azure's Big Data guide → <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing#challenges>



STREAMING DATA: BUSINESS IMPLICATIONS

1. Business decisions can potentially move at the speed of your customers' actions and quickly identify emerging trends.
 2. Business mindset must switch from “measure and process only what matters everyday” to “measure everything everyday and only process what matters today”
 3. As all industries digitize, gaps in streaming data capabilities means your competitors can move faster to take advantage of emerging trends.
- 
- 
- 

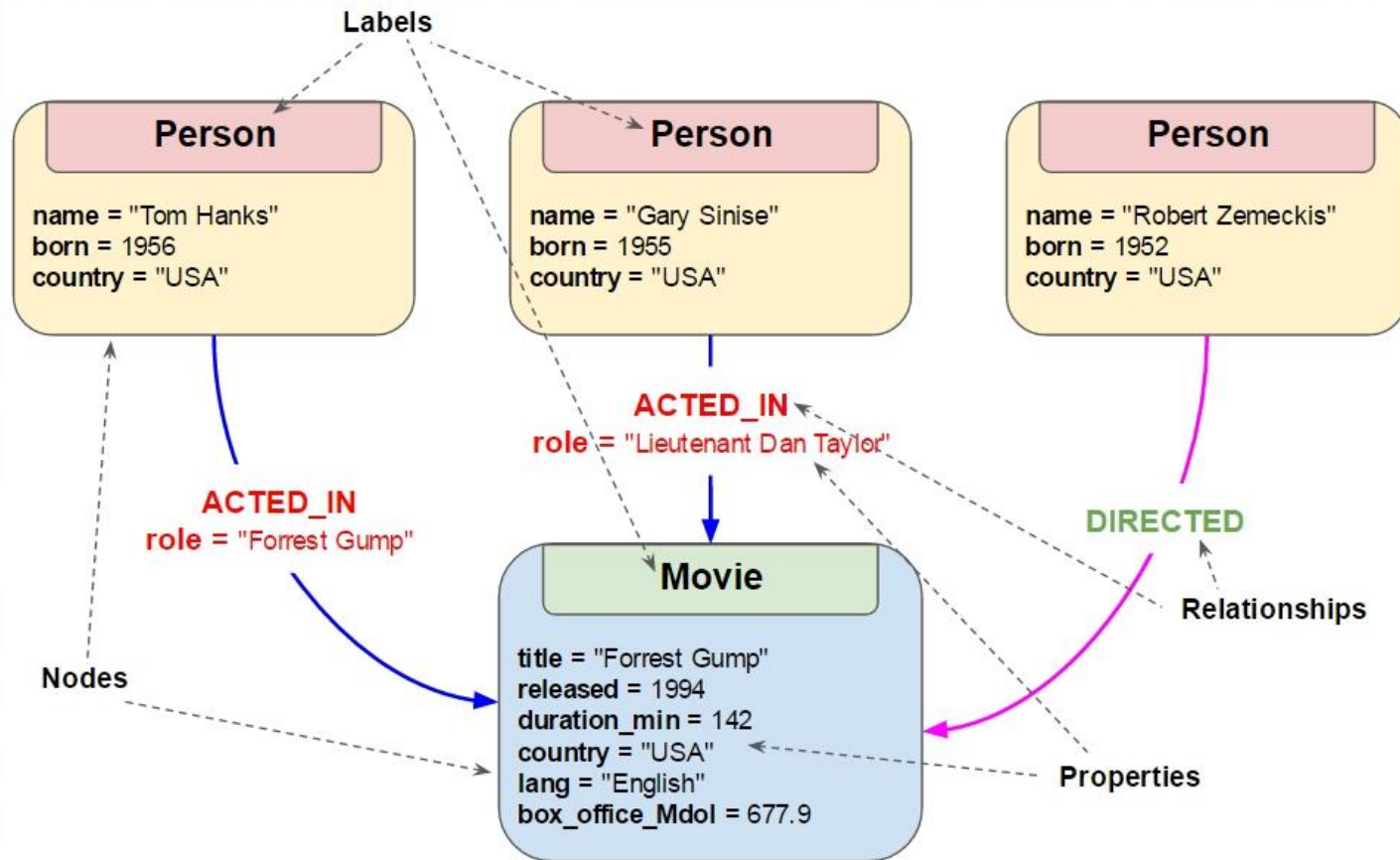
NOSQL

- “Not Only” SQL database structures that go beyond the relationship-based data designs we built using Entity-Relationship concepts
- Emerged during the 2000s to solve problems created by the collision of big data and SQL-based limitations:
 - Expanding database scale to process more and more data was expensive and error-prone
 - Expanding database scope to handle new data types and relationships was expensive and slow

NOSQL DATA STORAGE TYPES

- **Graph stores** – store data as a graph/network of nodes and edges
- **Document databases** – store data as documents with a flexible structure
- **Key-value stores** – store data as keys to reference one or more values
- **Search / wide-column stores** – optimized data storage for searching across semi-structured log data

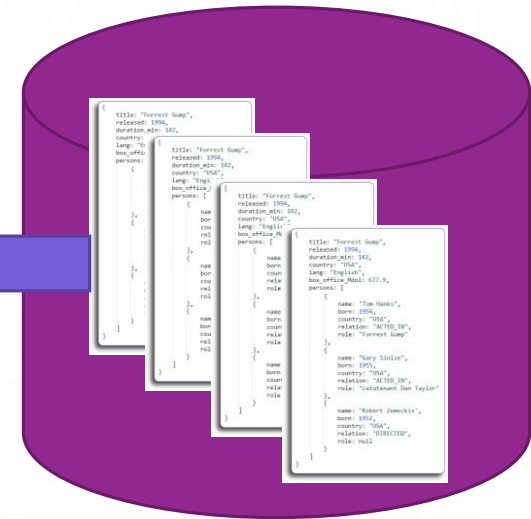
NOSQL: GRAPH DATA



NOSQL: DOCUMENT DATA

```
{
  title: "Forrest Gump",
  released: 1994,
  duration_min: 142,
  country: "USA",
  lang: "English",
  box_office_Mdol: 677.9,
  persons: [
    {
      name: "Tom Hanks",
      born: 1956,
      country: "USA",
      relation: "ACTED_IN",
      role: "Forrest Gump"
    },
    {
      name: "Gary Sinise",
      born: 1955,
      country: "USA",
      relation: "ACTED_IN",
      role: "Lieutenant Dan Taylor"
    },
    {
      name: "Robert Zemeckis",
      born: 1952,
      country: "USA",
      relation: "DIRECTED",
      role: null
    }
  ]
}
```

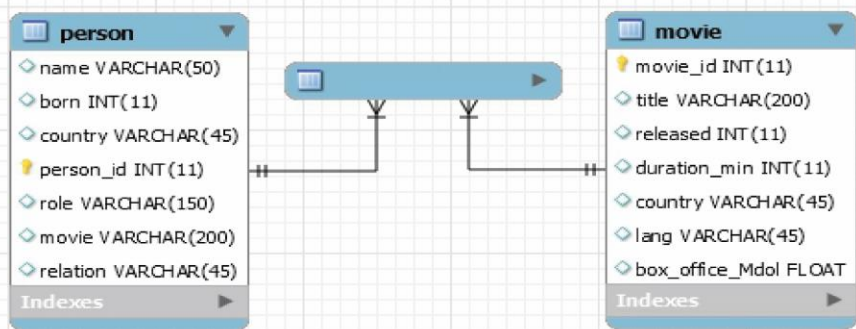
DOCUMENT



COLLECTION
(of Documents)

NOSQL DOCUMENT VS SQL

SQL's Rows within Tables and Relationships



	person_id	name	born	country	relation	role	movie
▶	0	Tom Hanks	1956	USA	ACTED_IN	Forrest Gump	Forrest Gump
	1	Gary Sinise	1955	USA	ACTED_IN	Lieutenant Dan Taylor	Forrest Gump
	2	Robert Zemeckis	1952	USA	DIRECTED		Forrest Gump

	movie_id	title	released	duration_min	country	lang	box_office_Mdol
▶	0	Forrest Gump	1994	142	USA	English	677.9



NOSQL's Document with Nested Documents

```
{
  title: "Forrest Gump",
  released: 1994,
  duration_min: 142,
  country: "USA",
  lang: "English",
  box_office_Mdol: 677.9,
  persons: [
    {
      name: "Tom Hanks",
      born: 1956,
      country: "USA",
      relation: "ACTED_IN",
      role: "Forrest Gump"
    },
    {
      name: "Gary Sinise",
      born: 1955,
      country: "USA",
      relation: "ACTED_IN",
      role: "Lieutenant Dan Taylor"
    },
    {
      name: "Robert Zemeckis",
      born: 1952,
      country: "USA",
      relation: "DIRECTED",
      role: null
    }
  ]
}
```

Amazon expert explains (00:00 to 02:36):

<https://www.youtube.com/watch?v=MBODF1Vru2Y>

SQL VS NOSQL: TERMINOLOGY

SQL	MongoDB	DynamoDB
Table	Collection	Table
Row	Document	Item
Column	Field	Attribute
Primary key	ObjectId	Primary key

NOSQL: KEY-VALUE DATA

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

EXAMPLE: SNAPCHAT STORIES (03:14 to 07:31) →

https://www.youtube.com/watch?reload=9&v=WUleQzu9I_8

NOSQL: BUSINESS IMPLICATIONS

- Business tech teams can deliver new software features faster to market
- Businesses can realize faster data performance and reduce data scaling costs
- Businesses can choose when an application's data adheres to Atomicity, Consistency, Isolation, and Durability constraints
- Businesses realize some increased application development costs as tech teams must implement custom APIs to work with NoSQL data, instead of relying on off-the-shelf SQL coding and processing engines

SQL: ACID STILL IMPORTANT FOR BUSINESS

- SQL is still useful for businesses as SQL systems guarantee data quality out-of-the-box by adhering to ACID, while NoSQL requires custom effort to do this.
- Atomicity → a transaction (add, remove, change) must execute completely or not at all.
- Consistency → every committed transaction matches the database schema.
- Isolation → each transaction executes separately from other transactions.
- Durability → the database can always recover from a during-transaction system failure to the last known state.
- Key applications where SQL continues to be important because of ACID: financials, enterprise resource planning

BIG DATA BASICS REFERENCES, PT 1

- BIG DATA

- IDC DataAge 2025 report: <https://www.seagate.com/our-story/data-age-2025/> and related coverage: <https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/>
- Ward & Barker's arXiv paper: <https://arxiv.org/abs/1309.5821>
- MIT's Technology Review on the history of the term: <https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/>
- Gartner's official definition: <https://www.gartner.com/it-glossary/big-data/>

- HADOOP/MAPREDUCE

- NYU High Performance Computing Lab MapReduce tutorial: <https://wikis.nyu.edu/display/NYUHPC/Big+Data+Tutorial+1%3A+MapReduce>
- Apache's MapReduce Tutorial: <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>

BIG DATA BASICS REFERENCES, PT 2

- NoSQL:
 - MongoDB tutorial (early pioneer): <https://www.mongodb.com/nosql-explained>
 - Amazon AWS tutorial (early adopter): <https://aws.amazon.com/nosql/>
- STREAMING/BATCH DATA
 - Microsoft Azure 's Batch Data processing solutions: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/batch-processing>
 - Microsoft Azure's Streaming Data process solutions: <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing>
 - O'Reilly's Streaming 101 and 102 blog posts by Tyler Akidau: <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101> and <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102>
 - "Streaming Systems" by Tyler Akidau, Slava Chernyak, and Reuven Lax. O'Reilly, 2018. ISBN-13: 978-1-491-98387-4.