BigData and NoSQL Overview

Outline

- What is BigData?
- What is NoSQL?
- NoSQL Database Types
- CAP Theorem
- BASE Principle
- ACID vs BASE
- RDBMS vs NoSQL

What is Big Data?

- Huge Amount of Data (Terabytes or Petabytes)
- Big data is the term for a collection of data sets so **large and complex** that it becomes **difficult** to process using on-hand database management tools or traditional data processing applications
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization

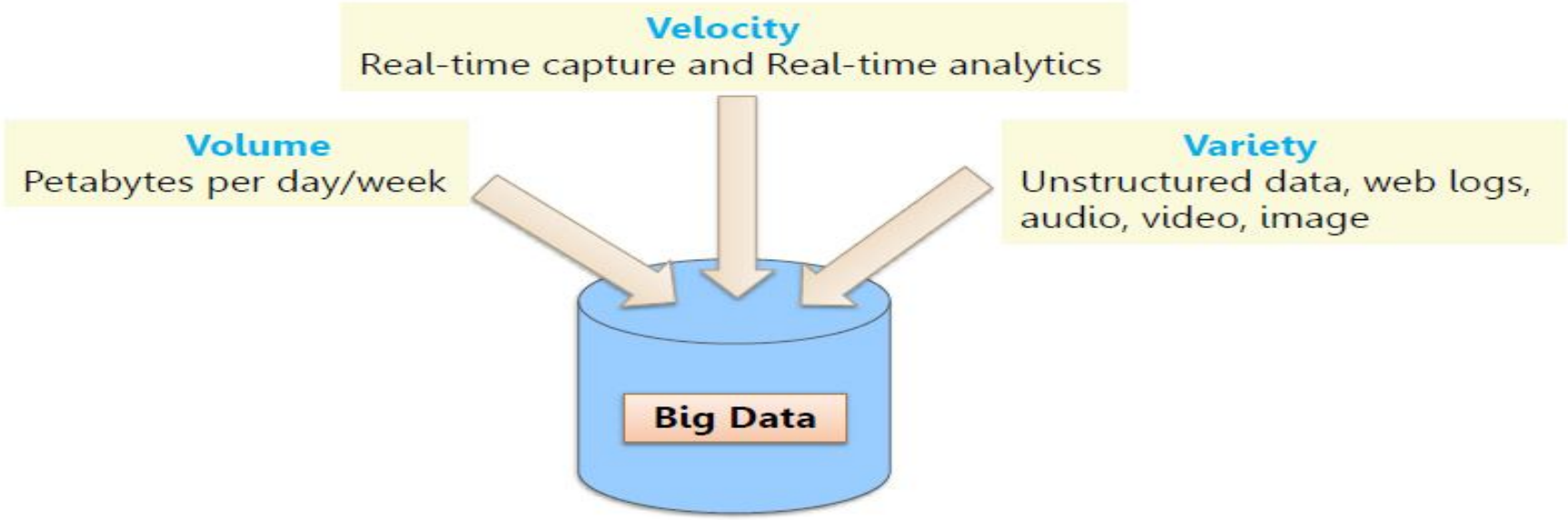
Sources of Big Data Generation

- Systems / Enterprises generate huge amount of data from Terabytes to Petabytes of information

NYSE generates about one terabyte of new trade data per day to perform stock trading analytics to determine trends for optimal trades

IBM's Definition of Big Data

- IBM's Definition – Big Data Characteristics
- http://www-01.ibm.com/software/data/bigdata/



Big Data in Industry

Web and e-tailing

- Recommendation Engines
- Ad Targeting
- Search Quality
- Abuse and Click Fraud Detection

Telecommunications

- Customer Churn Prevention
- Network Performance Optimization
- Calling Data Record (CDR) Analysis
- Analyzing Network to Predict Failure

Government

- Fraud Detection and Cyber Security
- Welfare Schemes
- Justice

Healthcare and Life Sciences

- Health Information Exchange
- Gene Sequencing
- Serialization
- Healthcare Service Quality Improvements
- Drug Safety

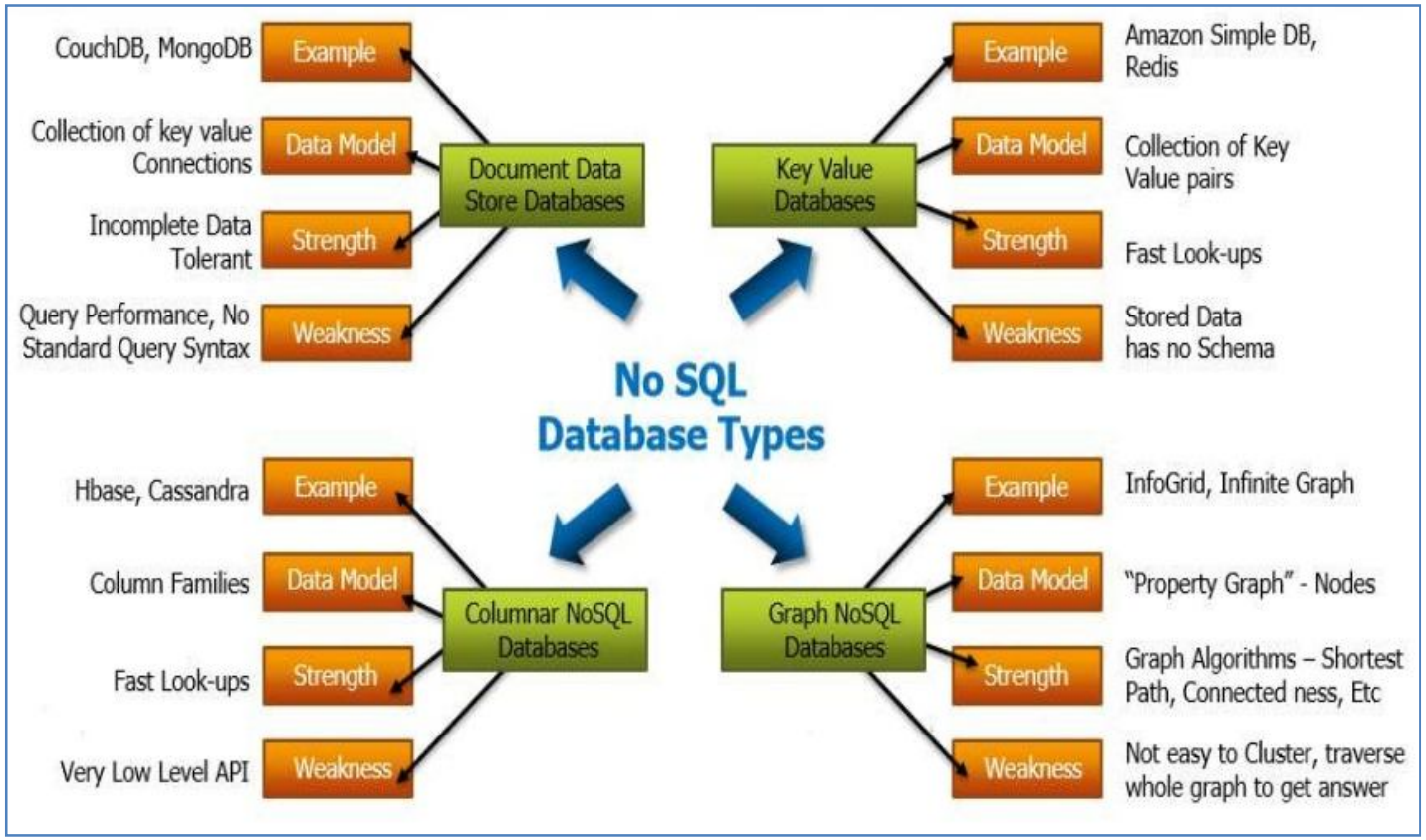
What is NoSQL?

- Stands for Not Only SQL
- Term was coined by Carlo Strozzi. Later redefined by Eric Evans
- Class of non-relational data storage systems
- Promotes schema less data model
- Encourages de-normalized data design
- Limited or No Joins
- Relaxation for one or more ACID properties

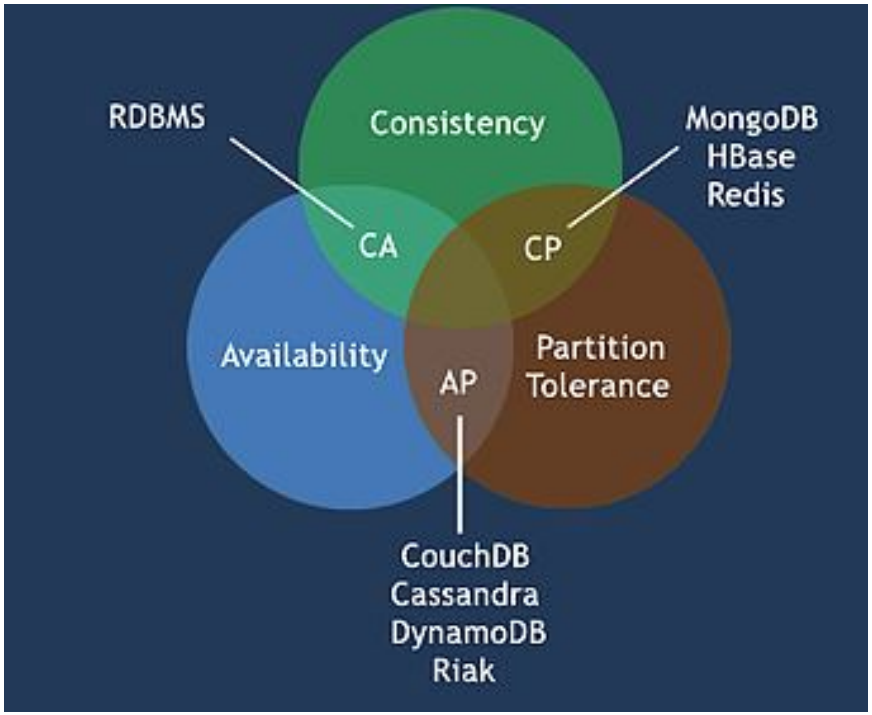
Key Features of NoSQL Database

NoSQL Database Types

NoSQL Database Types (cont..)



Brewer's CAP Theorem



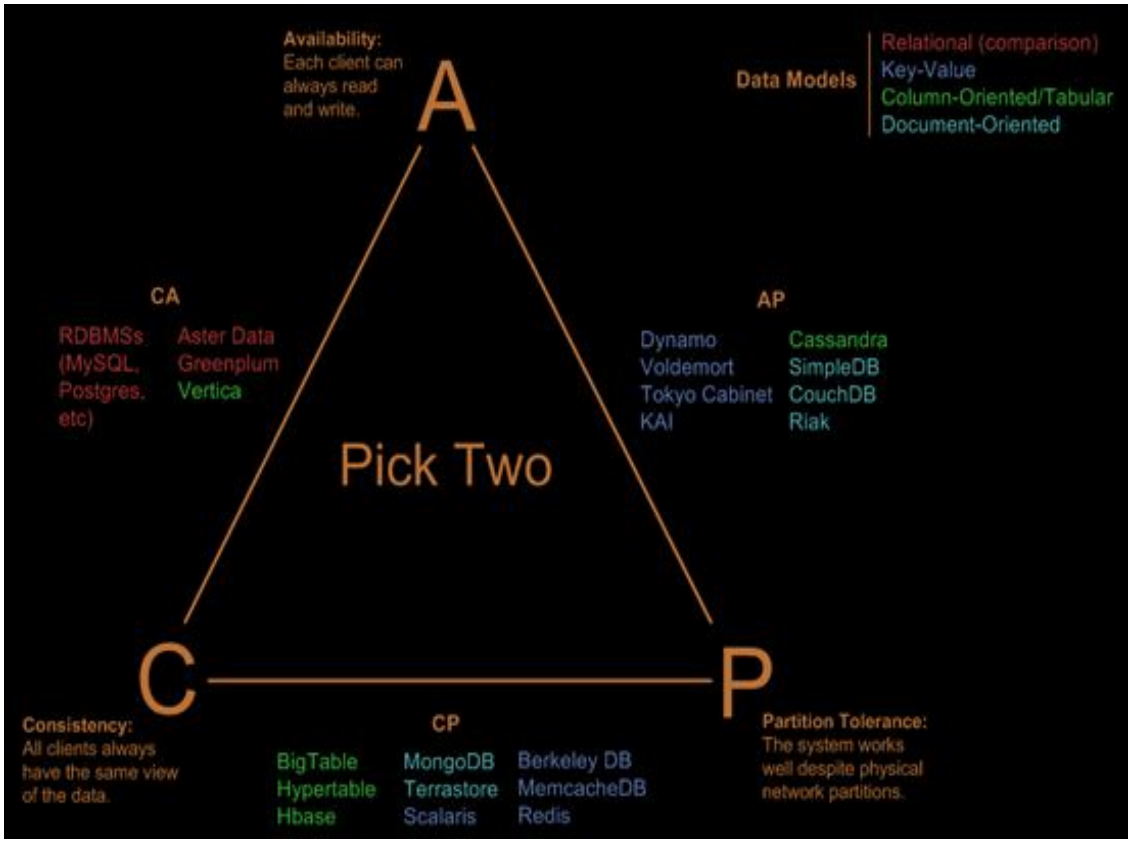
- Proposed by Eric Brewer, a computer scientist in 1999
- CAP is the basis for many NoSQL databases
- Brewer stated that its not possible to guarantee all three aspect. Any two will be picked and other one be compromised.

Consistency - all nodes in the cluster view the same data at the same time.

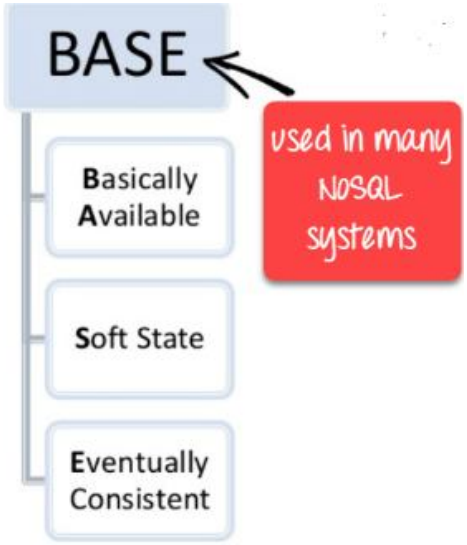
Availability - response is guaranteed for every request received (success or failure)

Partition tolerance - system continues to operate, despite the ad hoc message loss or failure on the part of the system.

Brewer's CAP Theorem (cont..)

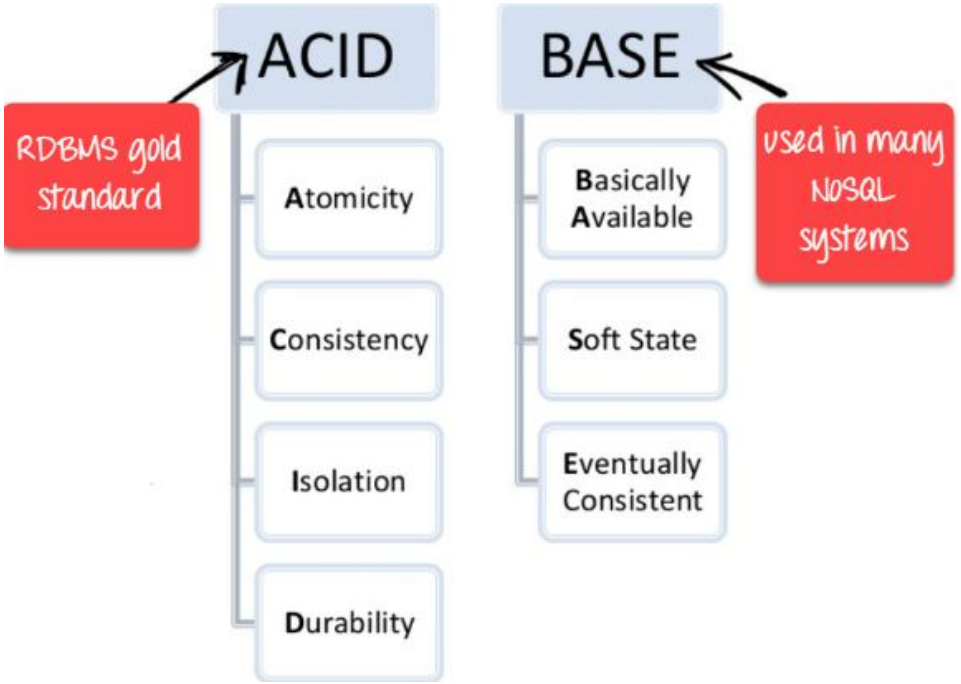


BASE Principle



- Basic Availability
  - Possibilities of faults but not a fault of the whole system.
- Soft-state
  - Copies of a data item may be inconsistent
- Eventually consistent
  - Copies becomes consistent at some later time if there are no more updates to that data item

ACID vs BASE



ACID	BASE
Strong consistency	Weak consistency – stale data OK
Isolation	Availability first
Focus on "commit"	Best effort
Nested transactions	Approximate answers OK
Availability?	Aggressive (optimistic)
Conservative (pessimistic)	Simpler!
Difficult evolution (e. g. schema)	Faster
	Easier evolution

RDMS vs NoSQL

RDBMS	NoSQL
• Structured and organized data • Structured Query Language (SQL) • Data and its relationships stored in separate tables. • Data Manipulation Language, Data Definition Language • Tight Consistency • ACID Transaction	• No declarative query language • No predefined schema • Key-Value pair storage, Column Store, Document Store, Graph Databases • Eventual consistency rather ACID property • Unstructured and unpredictable data • CAP Theorem • Prioritize high performance, high availability and scalability

Thank You