**Big Data Technology**

**(NoSQL and Hadoop)**

NoSQL(NOT ONLY SQL)

It is a light weight ,open source, non relational database that did not expose the standard SQL interface.

NoSQL databases are widely used in bigdata and other real time web applications.

**Features of NoSQL**:

1.NoSQL databases are non-relational

2.Distributed

3.No support for ACID properties

4.No fixed table schema

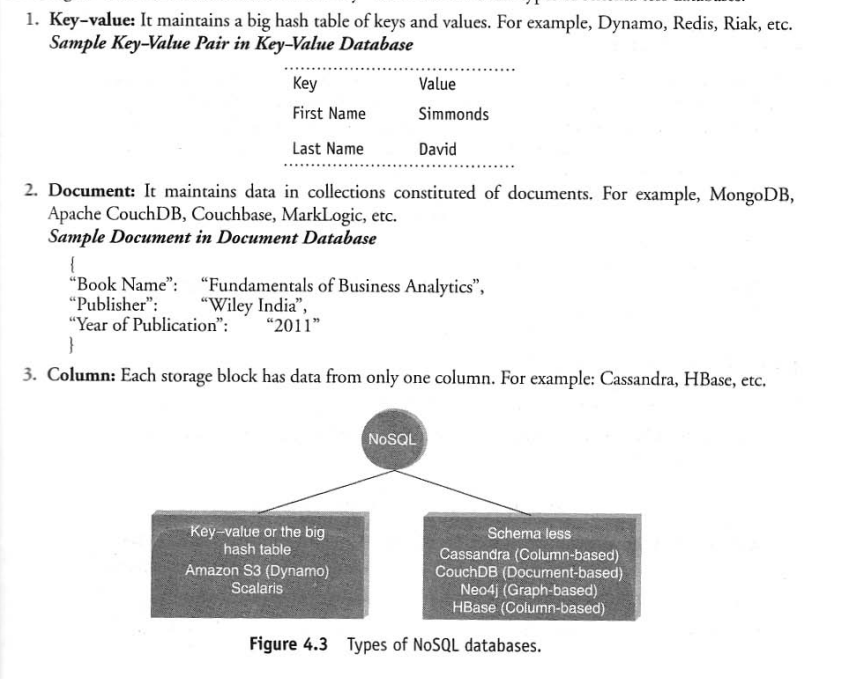
**Types of NoSQL Databases**

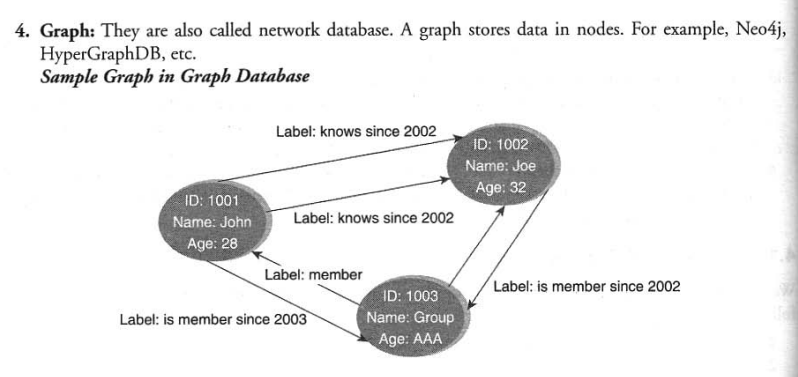
1. Key-value

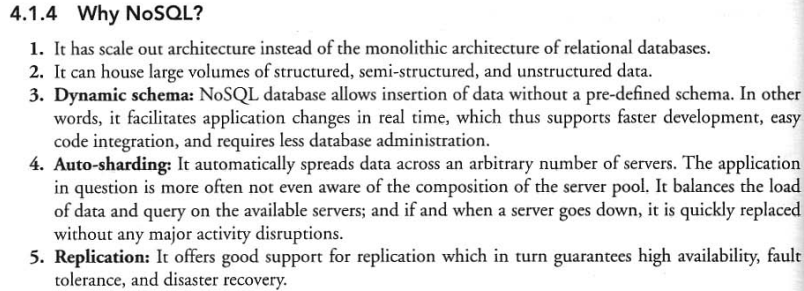
2. Document

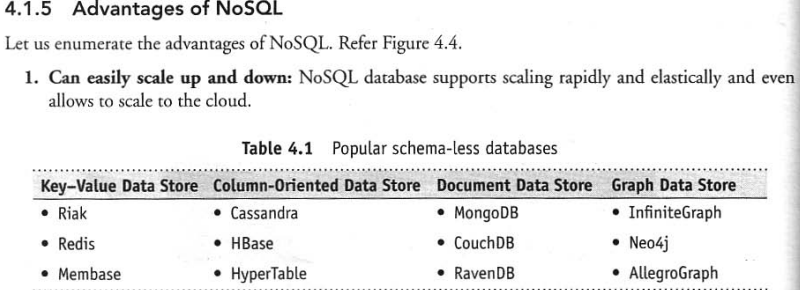
3. Column

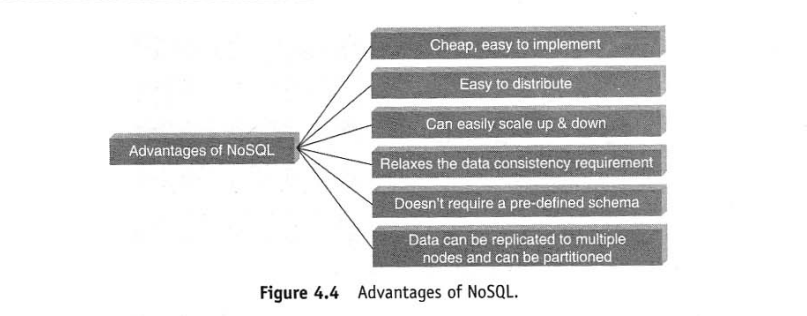
4. Graph

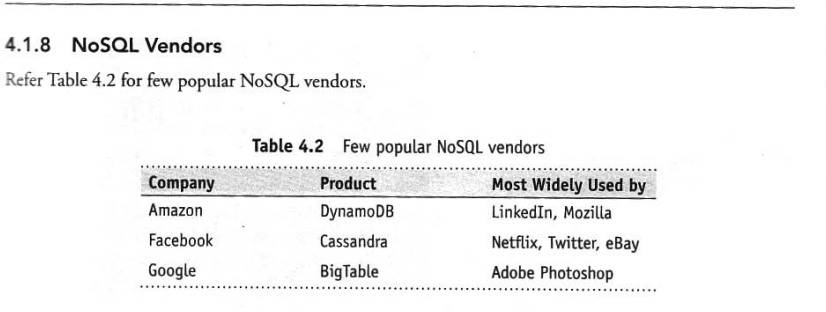


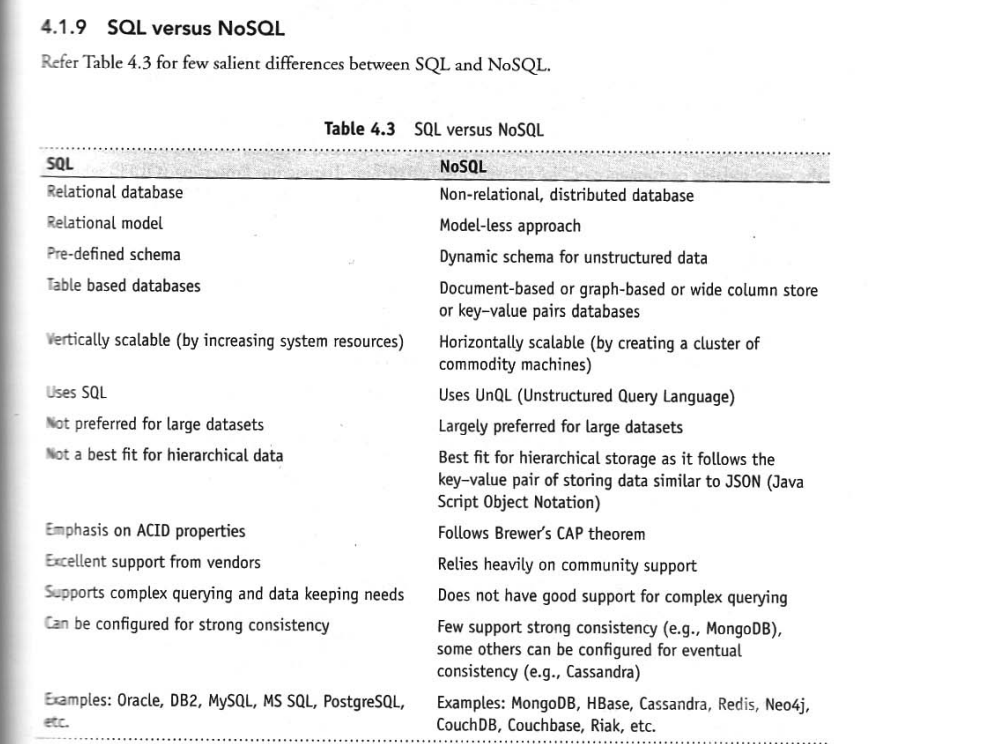












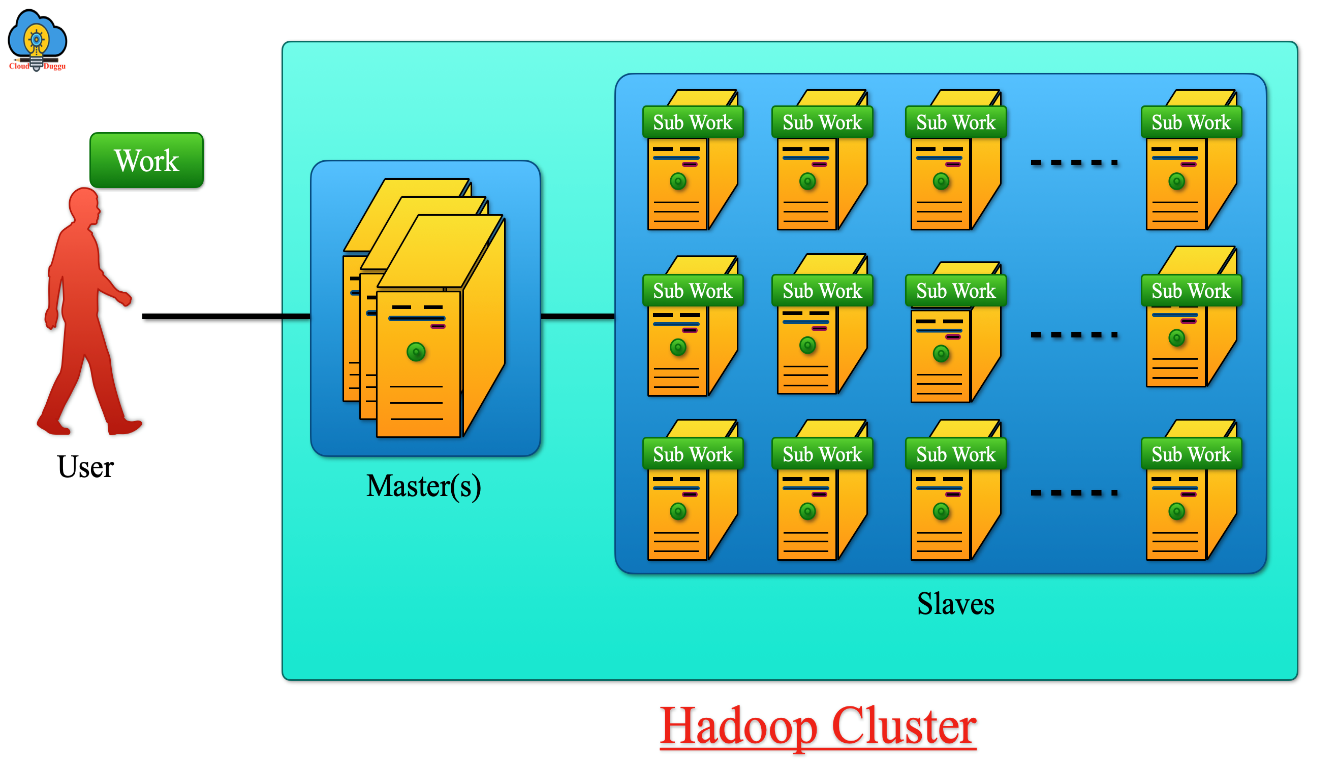
**Hadoop** is an open-sourceplatformforstorageandprocessingofdiversedatatypesthatenablesdata-drivenenterprisestorapidlyderivethecompletevaluefromalltheirdata.

**History of Hadoop**

The original creators of Hadoop are **Doug Cutting** (used to be atYahoo! now at Cloudera) and **Mike Cafarella** (now teaching at theUniversity of Michigan in Ann Arbor). Doug and Mike were buildinga project called **“Nutch**” with the goal of creating a large Web index.TheysawtheMapReduceandGFSpapersfromGoogle,whichwere obviously super relevant to the problem Nutch was trying tosolve. They integrated the concepts from MapReduce and GFS intoNutch;thenlaterthesetwocomponentswerepulledouttoformthegenesisoftheHadoopproject.

The name “Hadoop” itself comes from Doug’s son yellow plush elephant toy that he has.

* The scalability and elasticity of free, open-source Hadoop running onstandard hardware allow organizations to hold onto more data thaneverbefore.
* Hadoophandlesavarietyofworkloads,includingsearch,logprocess-ing, recommendation systems, data warehousing, and video/imageanalysis
* ApacheHadoopisanopen-sourceproject Hadoopisabletostoreanykindofdatainitsnativeformatandto perform a wide variety of analyses and transformations on that data.Hadoopstoresterabytes,andevenpetabytes,ofdatainexpensively.Itis robust and reliable and handles hardware and system failures auto-matically,withoutlosingdataor interruptingdataanalyses.
* **Hadoop runs on clusters of commodity servers and each of thoseservers has local CPUs and disk storage that can be leveraged by thesystem.**



ThetwocriticalcomponentsofHadoopare:

* 1. **TheHadoopDistributedFileSystem(HDFS)**.HDFSisthestorage system for a Hadoop cluster. When data lands in the cluster,HDFS breaks it into pieces and distributes those pieces among thedifferent servers participating in the cluster. Each server stores justa small fragment of the complete data set, and each piece of data isreplicatedonmorethanoneserver.

**2.MapReduce.** Because Hadoop stores the entire dataset in small piecesacross a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reportsits results back for collation into a comprehensive answer. MapReduceistheagentthatdistributestheworkandcollectstheresults.

* BothHDFSandMapReducearedesignedtocontinuetoworkinthefaceofsystemfailures.
* Because of the way that HDFS and MapReduce work,Hadoop provides scalable, reliable, and fault-tolerant services for data storageandanalysisatverylowcost.

**ComputeCluster**

Data

datadata data datadatadatadata data datadatadatadatadatadatadata

datadata data datadatadatadata data datadatadatadatadatadatadata

datadata data datadatadatadata data datadatadatadatadatadatadata

datadata data datadatadatadata data datadatadatadatadatadatadata

Map

DFSBlock1

DFSBlock2

Reduce

DFSBlock3

DFSBlock3

Map

p

Ma

DFSBloc

DFSBlock2

DFSBlock1

DFSBlock1

## Oldvs.NewApproaches

* Theoldwayisadataandanalyticstechnologystackwithdifferent layers “cross-communicating data” and working on “scale-up”expensive hardware.
* The new way is a data and analytics platformthat does all the data processing and analytics in one “layer,” with-out moving data back and forth.

Summary

1. The technology stack has changed. New proprietary technologiesand open-source inventions enable different approaches that make iteasierandmoreaffordable tostore,manage,andanalyzedata.
2. Hardwareandstorageisaffordableandcontinuingtogetcheapertoenablemassiveparallelprocessing.
3. The variety of data is on the rise and the ability to handle unstructureddataisontherise.

## DataDiscovery:WorktheWayPeople’sMindsWork

Tableau Software and QlikTechInternational.(Qlikview)

## Open-SourceTechnologyforBigDataAnalytics

* Open-source software is computer software that is available in source codeform under an open-source license that permits users to study, change, andimprove and at times also to distribute the software.
* Although the source code is released, there are still governing bodies andagreements in place. The most prominent and popular example is the GNU GeneralPublicLicense(GPL),which“allowsfreedistributionundertheconditionthatfurtherdevelopmentsandapplicationsareputunderthesamelicense.”Thisensuresthattheproductskeepimprovingovertimeforthegreaterpopulationofusers.
* Some other open-source projects are managed and supported by commercialcompanies,suchasCloudera,thatprovideextracapabilities,training,andprofessional services that support open-source projects such as Hadoop.
* You can make it into what you want and what you need. If you come up withan idea, you can put it to work immediately. That’s the advantage of the open-sourcestack—flexibility,extensibility,andlowercost.”
* “Oneofthegreatbenefitsofopensourceliesintheflexibilityoftheadoptionmodel:youdownloadanddeployitwhenyouneedit”.
* Pace ofsoftwaredevelopmenthasaccelerateddramaticallybecauseofopen-sourcesoftware.
* The old model was top-down, slow, inflexible and expensive. Thenew software development model is bottom-up, fast, flexible, andconsiderablylesscostly.
* A traditional proprietary stack is defined and controlled by asingle vendor, or by a small group of vendors. It reflects the old command-and-control mentality of the traditional corporate world andtheoldeconomicorder.
* An open-source stack is defined by its community of users and contributors. No one “controls” an open-source stack, and no one canpredict exactly how it will evolve. The open-source stack reflects thenew realities of the networked global economy, which is increasinglydependentonbigdata.

## TheCloudandBigData

With a cloud model, you pay on a subscription basis with no upfront capital expense. You don’t incur the typical 30 percent maintenance fees—and allthe updates on the platform are automatically available.

The ability to build massively scalable platforms—platforms where youhavetheoptiontokeepaddingnewproductsandservicesforzeroadditionalcost—is giving rise to business models that weren’t possible before.

**Key Features of the Cloud Model**

**1. Subscription-Based Model:**

* **No Upfront Capital Expense:**
  + Traditional IT infrastructure required businesses to invest heavily in hardware, software licenses, and data centers upfront.
  + With the cloud, costs are **operational (OPEX)** rather than **capital (CAPEX)**. Businesses pay on a subscription basis, often determined by usage (e.g., storage, compute power, bandwidth).
* **Flexibility and Predictability:**
  + Subscription models allow businesses to align costs with actual usage, enabling better financial planning and cost control.

**2. Lower Maintenance Costs:**

* **No Maintenance Fees:** Traditional systems often involve annual maintenance fees (~30% of the initial cost) for software updates, patches, and support.
* **Automatic Updates:** In cloud systems, providers handle updates, patches, and infrastructure upgrades automatically. This reduces downtime and ensures users always access the latest features and security enhancements without additional effort or cost.

**3. Scalability and Elasticity:**

* **Massive Scalability:**
  + Cloud platforms can handle exponential growth by dynamically scaling resources (e.g., adding more storage or compute power) based on demand.
  + This is critical for handling the large, fast-moving datasets typical of **Big Data**.
* **Cost Efficiency for Scaling:** Businesses can add new products and services without incurring significant additional costs, as cloud providers already have the infrastructure in place.

**Impact on Business Models**

The cloud model's unique features have enabled new business possibilities, especially in the context of **Big Data**:

**1. Pay-As-You-Go Innovation:**

* Businesses can experiment with new ideas and technologies without large initial investments. For example:
  + **Startups** can launch data-driven services without owning servers or data centers.
  + **Enterprises** can run short-term Big Data projects (e.g., analytics, AI/ML) without long-term commitments.

**2. Democratization of Technology:**

* Cloud services make advanced technology (like machine learning, data analytics, and IoT platforms) accessible to organizations of all sizes.
* Previously, only large companies with significant budgets could afford such capabilities.

**3. Business Agility and Faster Time-to-Market:**

* With scalable platforms and pre-built services, businesses can quickly develop and deploy new products, adapting to market changes rapidly.
* For example, streaming platforms can dynamically scale their infrastructure to handle peaks during global premieres without service interruption.

**4. Economies of Scale:**

* Cloud providers spread the cost of infrastructure across thousands of users, making advanced infrastructure affordable for individual businesses.
* For instance, a business doesn’t need to invest in high-performance servers to run **real-time analytics**; they can leverage cloud services like AWS Lambda or Google BigQuery.

**Role in Big Data**

Big Data refers to extremely large datasets that require advanced tools for processing and analysis. The cloud complements Big Data in several ways:

* **High-Performance Computing (HPC):** The cloud offers massive computational power on demand, enabling real-time processing of Big Data.
* **Data Storage:** Cloud platforms provide virtually unlimited storage capacity, making it feasible to store and manage petabytes or exabytes of data.
* **Data Integration and Collaboration:** Cloud-based Big Data platforms allow seamless integration of data from multiple sources and collaboration across geographically distributed teams.
* **Cost Efficiency:** By eliminating the need for dedicated hardware, the cloud lowers the entry barrier for businesses wanting to utilize Big Data analytics.

**Examples of Cloud-Enabled Business Models**

1. **Software-as-a-Service (SaaS):**
   * Tools like Salesforce, Tableau, or Microsoft Power BI are offered as subscription services, enabling businesses to access advanced capabilities without infrastructure investments.
2. **Data-Driven Services:**
   * Platforms like Spotify or Netflix utilize cloud infrastructure to store massive amounts of user data and deliver personalized experiences.
3. **E-Commerce and IoT:**
   * Businesses like Amazon and Tesla use scalable cloud platforms to process Big Data for inventory management, customer behavior analysis, and real-time IoT updates.