Cloud Computing

- In Wikipedia
  - Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure in the "cloud" that supports them.
- Related keywords
  - Key-Value Store, MapReduce, Hadoop

In Industry

SaaS: Software as a Service
PaaS: Platform as a Service
IaaS: Infrastructure as a Service

- Google: Gmail, AppEngine (Bigtable+Python)
- Yahoo!: PNUTS/Sherpa
- Amazon: S3, EC2
- Microsoft: Azure
- Sun: Sun Cloud
- IBM: XaaS
- …

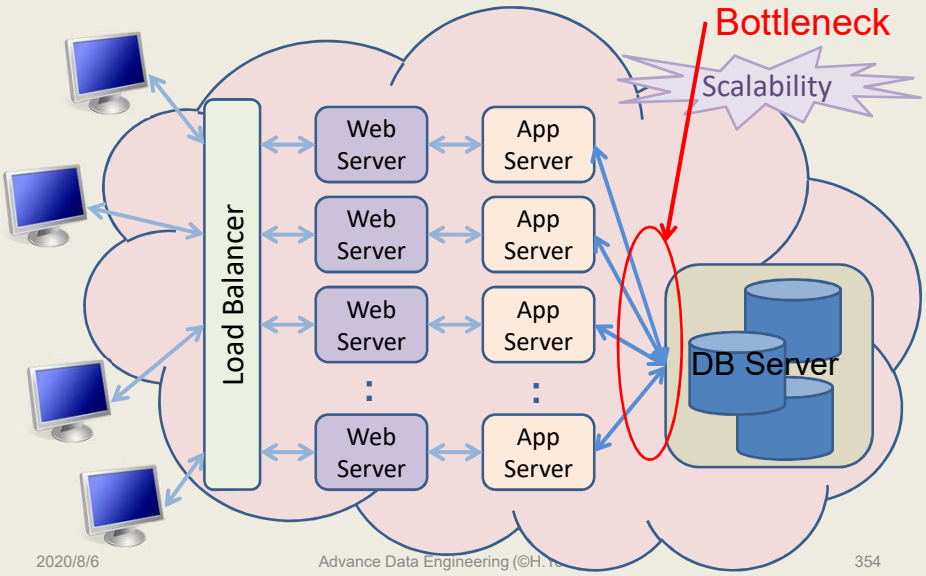
[Source http://markusklems.wordpress.com/]

Requirements to a Cloud Center

- Scalability
  - "Major Issue" of the Cloud Computing
    - in CACM 2008 article by Brian Hayes
  - Scale-out performance with keeping dependability
    - Including Security & Privacy
- Flexible configuration transparent to clients
  - Virtualization
- Low energy consumption

- Hardware vs. Software
  - Of course hardware is important
  - But, software (especially database technology) plays more important role to satisfy them

Bottleneck of the N-tier Model



Available Software Related to KVS

- Apache Hadoop
  - Consist of a number of sub-projects
    - Include Hbase and Hive
- Apache Cassandra
  - Amazon's Dynamo + Google's Bigtable
    - Dynamo: Key-Value Store
    - Bigtable: ColumnFamily-base data model
- Voldemort
  - Distributed Key-Value Storage System
- …

Key-Value Store (KVS)

- Simple data model
  - Pairs of Key and Value (or Value List)
    - e.g. <car, 1>, <car, [1,1,1]>
      - car is key and 1 or [1,1,1] is value
- Main goal: Scalability
  - Does not support complex queries or aggregation
  - Neither supports strict consistency
    - Eventually consistency
- Applications: do not require ACID properties
  - Such as Google's search engine

MR Approach vs. Parallel DBMSs

- MapReduce and Parallel DBMSs: Friend or Foes?
  - Michael Stonebraker, Daniel Abadi, Devid J. Dewitt, Sam Madden, Erik Paulson, Andrew Pavlo, and Alexander Rasin
  - CACM, Vol. 53, No. 1, pp.64-71, Jan. 2010.

MapReduce → Cloud Approach

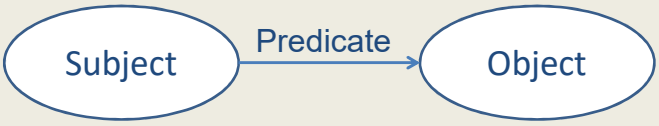
- Originally
  - A Comparison of Approaches to Large-Scale Data Analysis
  - Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Adabi, David J. Dewitt, Samuel Madden, Michael Stonebraker
  - SIGMOD'09, pp. 165-178, July 2009.

Differences

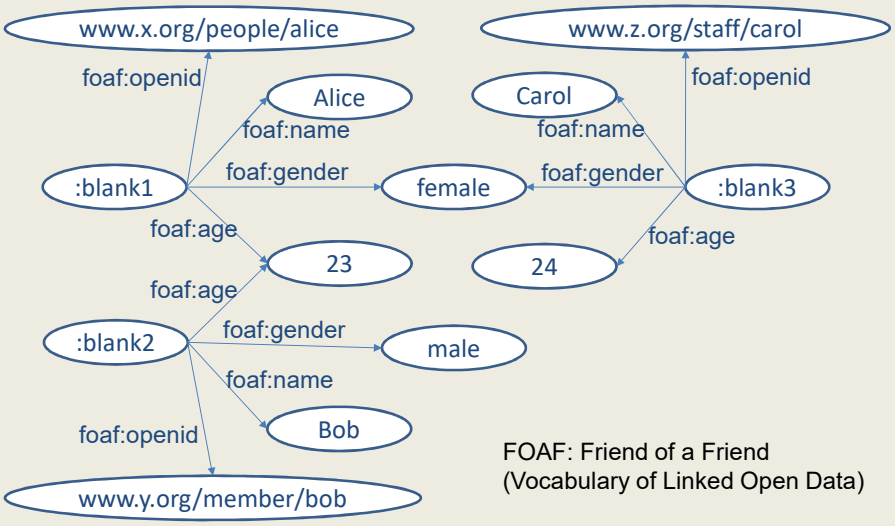
- MapReduce Approaches
  - Simple!
  - Suite for batch processes
    - Such as Extract-Transform-Load (ETL) system
- Parallel DBMS
  - Interactive query and update
  - Well optimized for structured information
- Conclusion
  - Complementary (not competitive)

RDF Data Model

- Triple: <Subject, Predicate, Object>
  - Subject: to indicate a resource (an entity)
  - Predicate: to represent a property of the entity
  - Object: a value of the property in form of a resource or literal
- A set of triples builds a directed graph
  - Subject and object are nodes
  - Predicate is a labeled directed edge from the subject to the object



An RDF Graph Example



FOAF: Friend of a Friend
(Vocabulary of Linked Open Data)

SPARQL Query Language

- Recursive definition (W3C)
  - SPARQL Protocol and RDF Query Language
- Example

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?x
WHERE {?y  foaf:name ?x
       ?y  foaf:age 23}
```

- Pattern match for <S, P, O>
- Self Join Operation

Graph Match by SPRQL

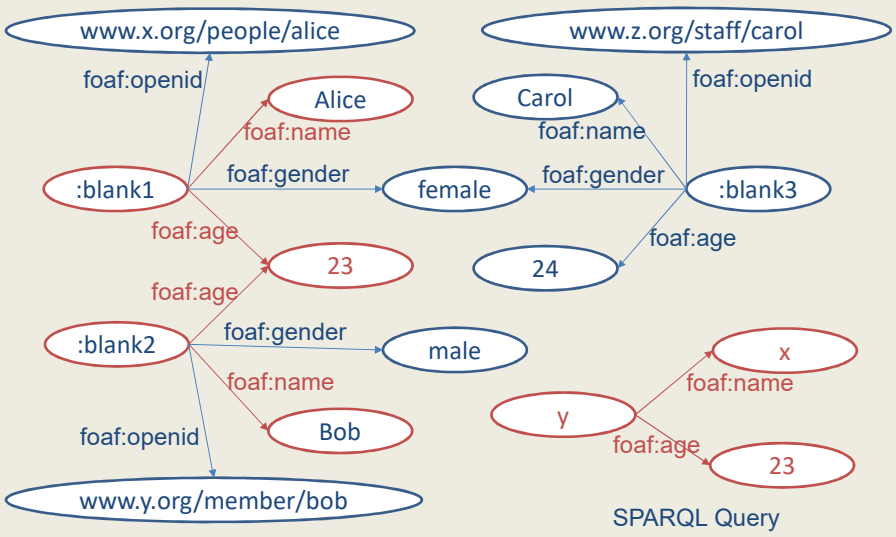


Image of Self Join

Subject	Predicate	Object
:blank1	foaf:name	Alice
:blank1	foaf:gender	female
:blank1	foaf:age	23
:blank1	foaf:openid	www.x.org/...
:blank2	foaf:name	Bob
:blank2	foaf:gender	male
:blank2	foaf:age	23
:blank2	foaf:openid	www.y.org/...
:blank3	foaf:name	Carol
:blank3	foaf:gender	female
:blank3	foaf:age	24
:blank3	foaf:openid	www.z.org/...

⋈

Subject	Predicate	Object
:blank1	foaf:name	Alice
:blank1	foaf:gender	female
:blank1	foaf:age	23
:blank1	foaf:openid	www.x.org/...
:blank2	foaf:name	Bob
:blank2	foaf:gender	male
:blank2	foaf:age	23
:blank2	foaf:openid	www.y.org/...
:blank3	foaf:name	Carol
:blank3	foaf:gender	female
:blank3	foaf:age	24
:blank3	foaf:openid	www.z.org/...

OWL (Web Ontology Language)

- A family of knowledge representation language as an extension of RDF vocabulary
  - OWL2 published by W3C OWL working group (2009)
    - OWL Lite, OWL DL (Description Logic), OWL Full
- RDF Schema (RDFS) is also ontology language
- Ex:

```
<owl:class rdf:ID="Mammal">
  <rdfs:subClassOf rdf:resource="#Animal"/>
  <owl:disjointWith rdf:resource="#Reptile"/>
</owl:Class>
```

Trends

- NoSQL (Not only SQL)
  - Note: Easy to misread, it does not deny SQL itself
- RDB: SQL
- OODB (Object-Oriented DB) : OQL (-> SQL)
  - ORDB (Object-Relational DB): PostgreSQL
- XML-DB: XQuery
- RDF: SPARQL (OWL/RDFS)
- KVS

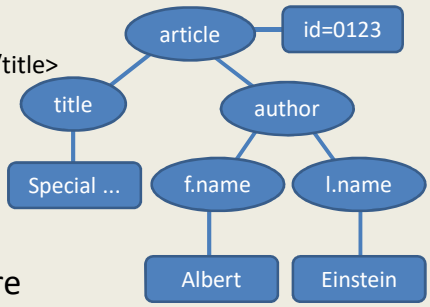
XML

- Extensible Markup Language
  - Origin: SGML (Standard Generalized Markup Language) 1986
  - Similar to HTML (Hyper Text Markup Language) for Web
  - XML 1.0: produced by W3C (World Wide Web Consortium) : 1998
- Trends
  - Recently, more and more data are represented and exchanged by XML on the Internet
    - A format for B2B data exchange
  - Store XML data as Database: XML-Database
    - Examples
      - Bibliography data: DBLP, ACM SIGMOD Record,
      - Bioinformatics data: TrEMBL, Swiss-Prot
      - …

Basic Syntax of XML

- Enclosed with <start tag> and <end tag>
  - <title>Special Theory of Relativity</title>
- Hierarchical Configuration (with attributes)
  - NG: <author>Einstein<title></author>Relativity</title>
  - Well formed :

```
<article id="0123">
  <title>Special Theory of Relativity</title>
  <author>
    <f.name>Albert</f.name>
    <l.name>Einstein</l.name>
  </author>
</article>
```



- Correspond to a tree structure

DTD

- Document Type Definition
  - To create rules for the elements in XML documents
- An example

```
<!DOCTYPE journal-db [
  <!ELEMENT journal (article)+>
  <!ELEMENT article (title, authors)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT authors (author)+>
  <!ELEMENT author (f.name, l.name)>
  <!ELEMENT f.name (#PCDATA)>
  <!ELEMENT l.name (#PCDATA)>
  <!ATTLIST article id IDREFS #REQUIRED>
]>
```

XML-DB vs. RDB

- RDB:
  - Fixed Schema : Table
    - Every tuple has all attributes
  - First Normal Form
    - Each element is atomic
  - Most common query language: SQL
- XML-DB
  - Flexible Schema: semi-structured
  - Hierarchical Configuration
  - Query langate: Xpath, XQuery

XPath

- A language to navigate, and find data, within XML documents.
  - XPath 1.0 became a Recommendation 1999 by W3C
  - XPath 2.0 is the current version
- Select one or more nodes to retrieve the data they contain
  - Expressed by location step
    - axis :: node-test [predicate]
  - axis: child, descendant, parent, following-sibling, …
  - /child::journal/child::article/child::authors/child::author/child::l. name
- Simplified
  - /journal/article/authors/author/l.name
  - /journal//l.name[Einstein]
  - //article[@id="0123"]

XQuery

- A language for querying XML data
  - Recommended by W3C
  - The means to extract and manipulate data from XML documents
    - To XML what SQL is to database tables
  - Use path expressions
  - Provides new XML documents to be constructed
- Syntax
  - FLWOR
    - FOR, LET, WHERE, ORDER BY, RETURN

Example of FLWOR of XQuery

```
for $x in doc("journal-db.xml")//article
let $a := $x/authors/author
where $x/year < 1930
order by $x//l.name descending
return
<articles-before-1930>
  <title>{$x/title}</title>
  <author>{$a/l.name, $a/f.name}</author>
</articles-before-1930>
```

Answer Example for the Query

```
<articles-before-1930>
  <title>Special Theory of Relativity</title>
  <author>Einstein, Albert</author>
</articles-before-1930>
<articles-before-1930>
  <title>Uncertainty Principle</title>
  <author>Heizenberg, Werner</author>
</articles-before-1930>
```


Summary of this course

- Data Engineering: A research area managing Data
  - How to store, search for, retrieve and process data efficiently
- Data Warehousing
  - Stock up data of operational transaction processing
  - Apply OLAP (Data Cube) and data mining methods
- Devices to store a large amount data reliably
- Indexing methods to search for target data in storages
- Cost estimation of relational database operations
- Parallel relational database operations / Skew handling
- Distributed databases / B2B connections
- Data engineering for Cloud environment
  - HDFS, KVS, Semantic Web, Security, RDF DB, XML DB