Cloud computing

TOPICS

1)DATA AS A SERVICE

2)NOSQL AS A SERVICE

3)IDENTITY AS A SERVICE

What is DaaS ?

- Data stored in the cloud & provides on demand service to users regardless of their geographic location
- Cousin of SaaS
- Provides solution for data storage, integration, management and analytics
- A healthy combination of data science, strategy, and structure to make datasets understandable and actionable
- Two ways to use Data-as-a-Service: By outsourcing our own data, Taking advantages of public data managed by third party
- It involves tasks such as data looping, data integration, business integration, service oriented architecture (dynamic, highly configurable and collaborative app.) , data visualization, data mapping & enterprise search

- ***Traditional Approach vs. DaaS?***

1. Data as goods	1. Data as service
2. Bulky download	2. Dynamic access
3. Dated with time of download	3. Always latest update
4. Need for storage	4. Storage is provided
5. Complex access when a large amount	5. Easy and simple access of view
6. Analysts, researchers, enthusiasts	6. Dynamic content providers, Mashup creator

Benefits

- Data profitable – through data analysis, providers drive profits by delivering data
- Predictive analysis – Exploiting AI to efficiently interpret data to predict future behavior
- More Agile decision making – data centric decision (collaborative, iterative, and transparent)
- Cost-effectiveness – smaller staff requirements, dynamic allocation, processing costs easier to optimize
- Data quality – data driven culture, personalized customer experience, data visualization
- Faster / Easy access – Make data more accessible and digestible
- Larger storage, Large number of users
- Scalability / Flexibility – resources can be allocated instantaneously
- Reliability – lowered risks less likely to fail, workload less prone to downtime or disruption
- Maintenance- tools & services automatically managed & kept up-to-date by providers

Drawbacks

- Customers reliance on service providers' ability to avoid server downtime
- Generally data is not available for download

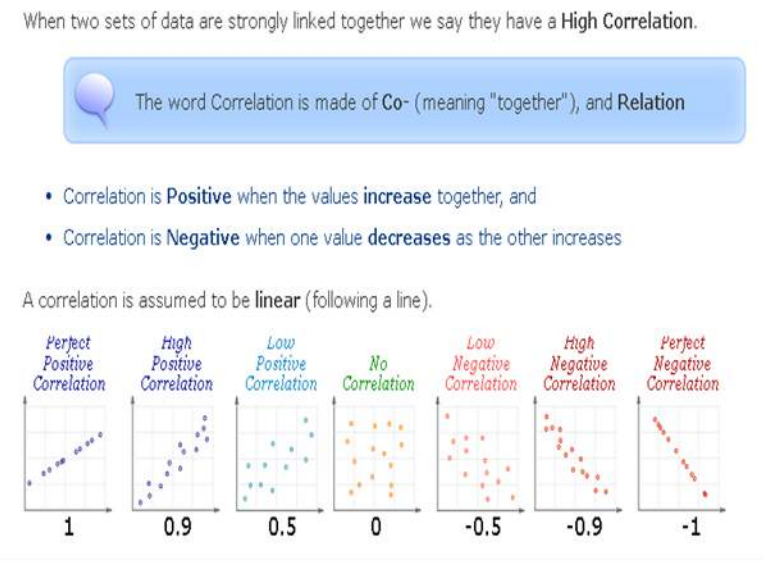
Challenges ?

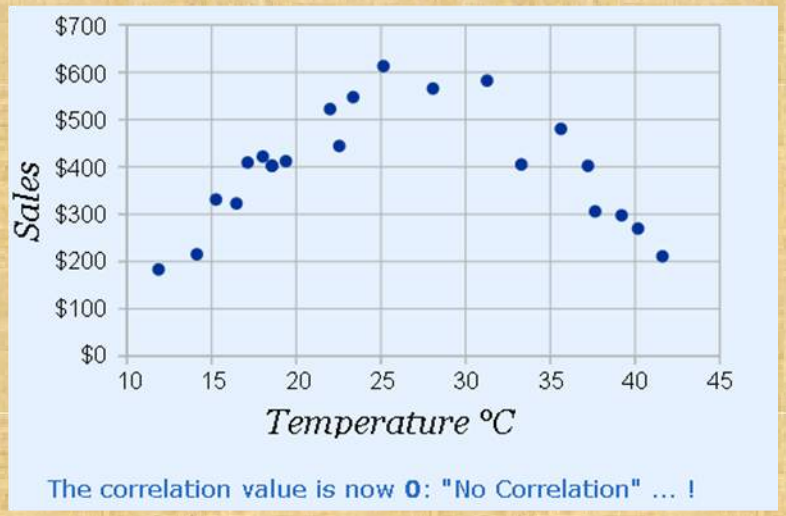
- Evaluate data silos (collection of information in an organization that is isolated from and not accessible by other parts of the organization)
- Privacy concerns (sharing data to other applications and services outside of the current application or departmental walls)
- Security concerns (Who can access the data, and how? Limiting access implies access control, which needs to be managed. If the data is going to be exchanged, especially between networks, will it be secure, and if so, how?)
- Falling short of true value (weighting of utility of free data)
- Data governance (Publishing and subscribing to data services require data governance to ensure the accuracy of the information being shared)

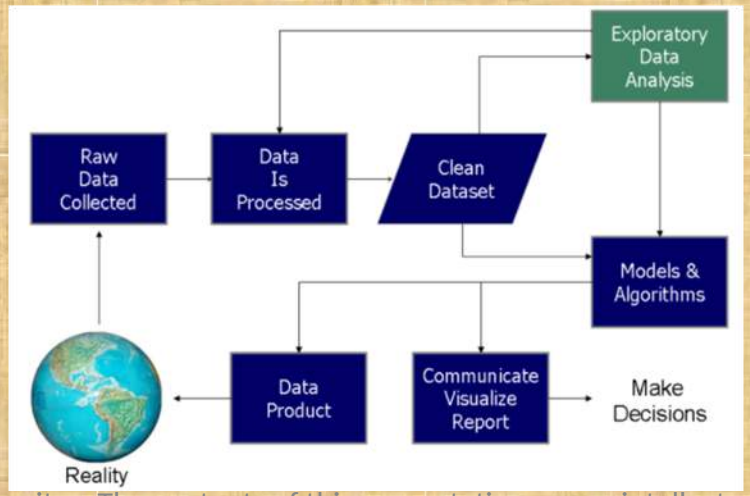
Key elements of DaaS

- Data collection – Identifying best and cleanest methodology & timing for gathering data & collecting insights
- Data aggregation – Process of compiling data points for a specific purpose, which are then analyzed & summarized into form of actionable insights.
- Data correlation – A statistical analysis looks at how strongly two points relates. Strong correlation shows strong relationship between two areas, allowing low risk decision making
- Statistical significance – A reliable method (Hypothesis testing) for measuring risk tolerance associated and confidence levels in datasets
- Data visualization– Identifying trends, patterns, or outliers and visually displaying insights in forms of graphs, tables, charts, maps and other visual formats, that enables companies to gain buy-in form teams and stakeholders, democratizing data makes it more accessible company-wide
- Advanced analytics– Process of developing complex models by simplifying big data with the goal of deepening insights (data mining, semantic/sentiment/graph/cluster/network analysis).

Key elements of DaaS





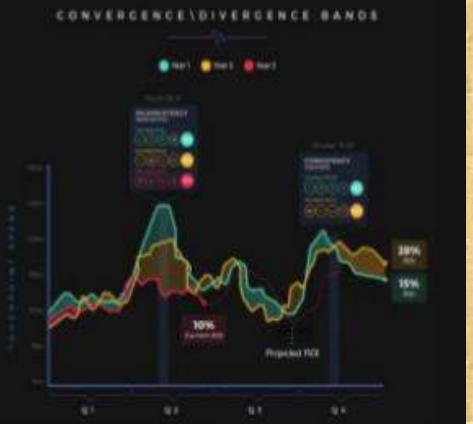


Key elements of DaaS

Making Visualized Data Insightful.

The way you visualize your data can determine the amount of buy-in you achieve and the depth of insight you deliver to your teams.

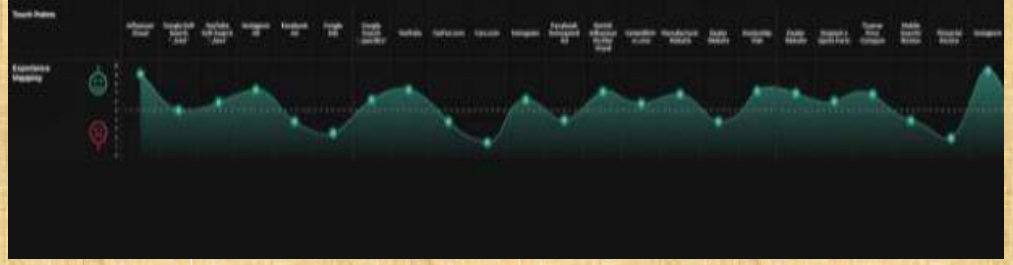
An effective DaaS provider will be able to visualize data that will show you where to spend strategically and how to see the biggest profits as a result of your data-centric initiatives.



Making Data Actionable.

Insights are only valuable if you put them into action and use the data at your disposal to make revenue-driving changes.

Data mapping is the process of making these insights actionable. By mapping out your next steps, you have a clear idea of where you must make strategic changes to see the biggest impact and ROI for your data-investment.



Data virtualization tool

- Red Hat Jboss: best choice for developers and those who are using micro services and containers
- TIBCO: helps administrators and users to create a data virtualization platform for accessing the multiple data sources and data sets and provides a built in transformation engine to combine non-relational and un-structured data sources
- Oracle data service integrator: mainly worked with Oracle products and allows organizations to quickly develop and manage data services to access a single view of data.
- SAS Federation Server: provides various technologies such as scalable, multi-user to access data from multiple data services and mainly focuses on securing data.
- Denodo: allows organizations to minimize the network traffic load and improve response time for large data sets., suitable for both small as well as large organizations.

What is NoSQL ?

- Stands for “Not Only SQL”
- A non-relational database (No Tables, No Predefined Schema)
- An umbrella term for all databases that don’t follow the RDBMS principles
- A collection of several (related) concepts about data storage and manipulation
- A Flexible database for Big data & Real-time Web apps
- Various NoSQL databases: Document (Mongo DB, Couch DB), Columnar (Apache Casendra, Hbase), Key-value pairs (Redis, Dynamo DB, Cauchbase, Graph-based (Neo 4J).
- From www.nosql-database.org:
- A next Generation, non-relational, distributed, open-source and horizontal-scalable Database Originally intended for modern web-scale databases.
- Schema-free, easy replication support, simple API, eventually consistent / BASE (not ACID), huge data amount, and more.

Where does NoSQL come from?

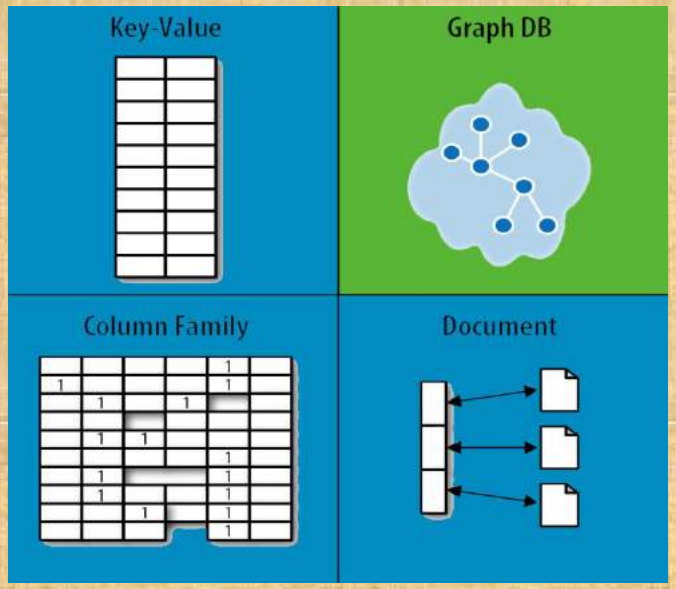
- Non-relational DBMSs are not new, but NoSQL is a new incarnation
  - Due to massively scalable Internet applications
  - Based on distributed and parallel computing
- Development
  - Starts with Google , first research paper published in 2003
  - Continues also thanks to Lucene's developers/Apache (Hadoop) and Amazon (DynamoDB) , then a lot of products and interests came from Facebook, Netflix, Yahoo, and many more
- Three major papers were the seeds of the NoSQL movement
  - Bigtable (Google), DynamoDB (Amazon), CAP Theorem
- Explosion of social media sites (Facebook, Twitter) with large data needs
- Rise of cloud-based solutions such as Amazon S3
- Just as moving to dynamically-typed languages (Python, Ruby, Groovy), a shift to dynamically-typed data with frequent schema changes
- Open-source community

NoSQL & Big data

- NoSQL comes from Internet, thus mostly related to the “big data” concept
- How much big are “big data”?
  - Over few terabytes Enough to start spanning multiple storage units
- Challenges
  - Efficiently storing and accessing large amounts of data is difficult, even more considering fault tolerance and backups
  - Manipulating large data sets involves running immensely parallel processes
  - Managing continuously *evolving schema* and metadata for *semi-structured and un-structured* data is difficult

NoSQL Database Types

Discussing NoSQL databases is complicated because there are a variety of types:



- **Sorted ordered Column Store**
  Optimized for queries over large datasets, and store
  columns of data together, instead of rows
- **Document databases**
  pair each key with a complex data structure known as a document.
- **Key-Value Store**
  are the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value.
- **Graph Databases**
  are used to store information about networks of data, such as social connections.

Document databases (Document store)

- Loosely structured sets of key/value pairs in documents, e.g., XML, JSON, BSON
- Encapsulate and encode data in some standard formats or encodings
- Documents treated as a whole, avoiding splitting a document into its constituent name/value pairs, addressed in the database via a unique key
- Database offers an API or query language that retrieves documents based on their contents.
- Documents are schema free, i.e., different documents can have structures and schema that differ from one another. (An RDBMS requires that each row contain the same columns.)
- Documents allow documents retrieving by keys or contents and Notable for **MongoDB (used in FourSquare, Github), CouchDB (used in Apple, BBC, Canonical, Cern)**

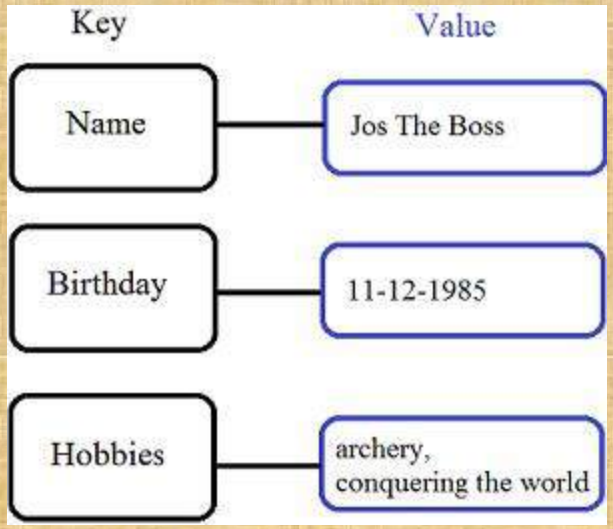
Syntax Example

```
{ _id: ObjectId("51156a1e056d6f966f268f81"),
  type: "Article",
  author: "Derick Rethans",
   title: "Introduction to Document Databases with MongoDB",
   date: ISODate("2013-04-24T16:26:31.911Z"),
   body: "This arti…“ },

{   _id: ObjectId("51156a1e056d6f966f268f82"),
  type: "Book",
  author: "Derick Rethans",
  title: "php|architect's Guide to Date and Time Programming with PHP",
  isbn: "978-0-9738621-5-7“ }
```

Key/Value stores

- Store data in a schema-less way
- Store data as maps
  - HashMaps or associative arrays
  - Provide a very efficient average running time algorithm for accessing data
- Notable for:
  - Couchbase (Zynga, Vimeo, NAVTEQ, ...)
  - Redis (Craiglist, Instagram, StackOverfow, flickr, ...)
  - Amazon Dynamo (Amazon, Elsevier, IMDb, ...)
  - Apache Cassandra (Facebook, Digg, Reddit, Twitter,...)
  - Voldemort (LinkedIn, eBay, ...)
  - Riak (Github, Comcast, Mochi, ...)



Sorted Ordered Column-Oriented Stores

- Data are stored in a column-oriented way
  - Data efficiently stored
  - Avoids consuming space for storing nulls
  - Columns are grouped in column-families
  - Data isn't stored as a single table but is stored by column families
  - Unit of data is a set of key/value pairs
    - Identified by "row-key"
    - Ordered and sorted based on row-key
- Notable for: Google's Bigtable (used in all Google's services), HBase (Facebook, StumbleUpon, Hulu, Yahoo!, ...)

Sales

Product	Customer	Date	Sale
Beer	Thomas	2011-11-25	2 GBP
Beer	Thomas	2011-11-25	2 GBP
Vodka	Thomas	2011-11-25	10 GBP
Whiskey	Christian	2011-11-25	5 GBP
Whiskey	Christian	2011-11-25	5 GBP
Vodka	Alexei	2011-11-25	10 GBP
Vodka	Alexei	2011-11-25	10 GBP

Product

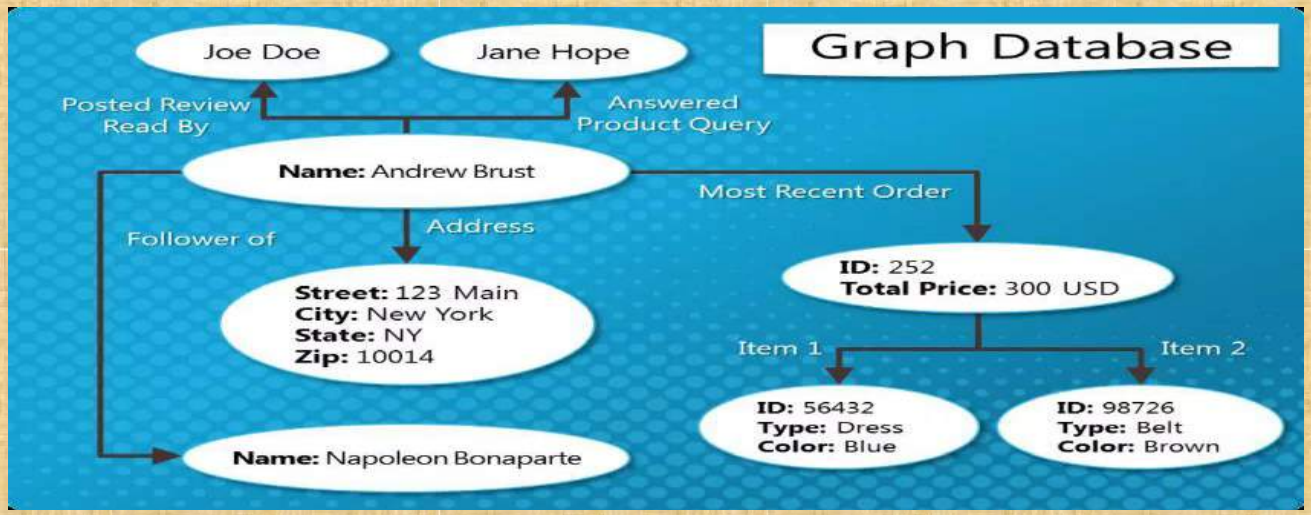
ID	Value
1	Beer
2	Beer
3	Vodka
4	Whiskey
5	Whiskey
6	Vodka
7	Vodka

Customer

ID	Customer
1	Thomas
2	Thomas
3	Thomas
4	Christian
5	Christian
6	Alexei
7	Alexei

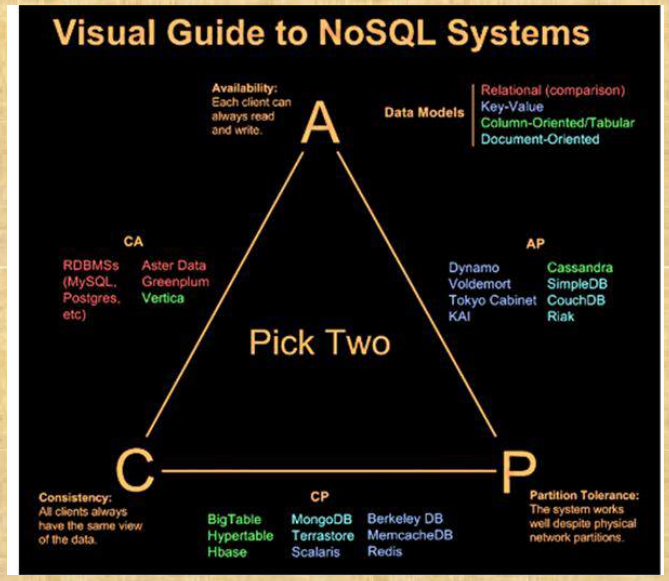
Graph Databases

- Graph-oriented
- Everything is stored as an edge, a node or an attribute.
- Each node and edge can have any number of attributes.
- Both the nodes and edges can be labelled.
- Labels can be used to narrow searches.



Characteristics of NoSQL

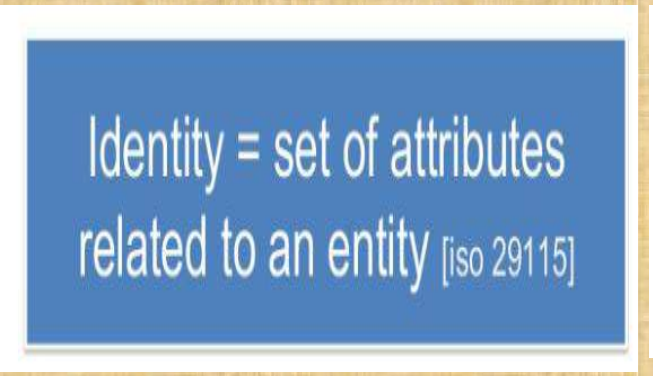
- Acronym contrived to be the opposite of ACID
- BASE: Basically Available, Soft state, Eventually Consistent
- Characteristics
- Weak consistency – stale data OK, Availability first, Best effort, Approximate answers OK, Aggressive (optimistic), Simpler and faster
- Follows CAP theorem
- At most two of the following three can be maximized at one time
- Consistency
  - Each client has the same view of the data
- Availability
  - Each client can always read and write
- Partition tolerance
  - System works well across distributed physical networks

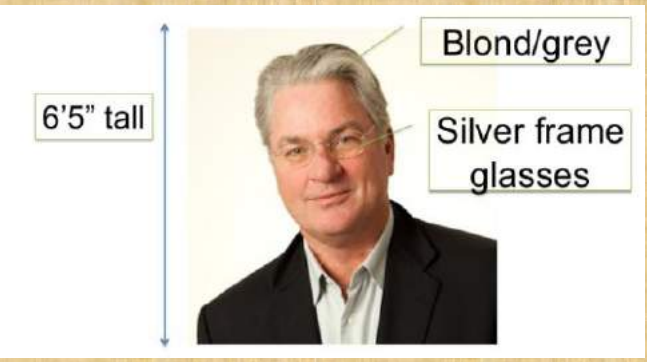


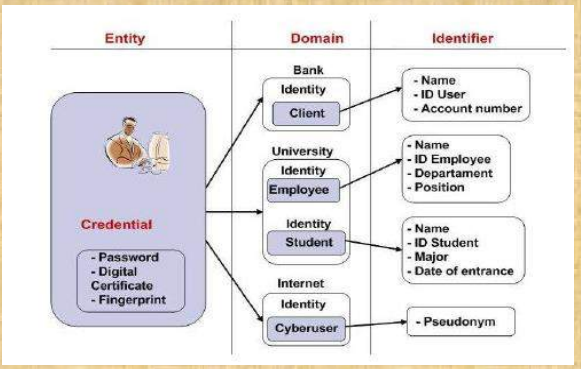
Performance

- There is no perfect NoSQL database
- Every database has its advantages and disadvantages
  - Depending on the type of tasks (and preferences) to accomplish
- NoSQL is a set of concepts, ideas, technologies, and software dealing with
  - Big data
  - Sparse un/semi-structured data
  - High horizontal scalability
  - Massive parallel processing
- Different applications, goals, targets, approaches need different NoSQL solutions
- Where would I use a NoSQL database?
- Do you have somewhere a large set of uncontrolled, unstructured, data that you are trying to fit into a RDBMS?
  - Log Analysis
  - Social Networking Feeds (many firms hooked in through Facebook or Twitter)
  - External feeds from partners
  - Data that is not easily analyzed in a RDBMS such as time-based data
  - Large data feeds that need to be massaged before entry into an RDBMS

What is IDaaS ?







- Attributes : biological, personal,, family & friends, belief & value, selection & choices, habits
- In computer network: Digital identity in terms of Single-factor authentication or Multi-factor authentication
- For user & machine, identity stored in domain security database, directory service, and in data store in federated system.
- Network interfaces identified by MAC addresses, S/w is identified by S/w product activation
- To validate websites, transactions & their participants, clients & network services, various forms of identification services have been deployed in the network E.g. Token providing service, Certificate server, other trust mechanism

Identity and access management Definition (IAM)

- A/c to Gartner, IAM is a security discipline that enables the right individuals to access the right resources at the right time for the right reasons

- Identity protection is one of more expensive and complex area of network computing. The migration of web applications to Cloud computing platform has raised concerns about the privacy of sensitive data belonging to the consumers of cloud services.

- How can consumers verify that a service provider conform to the privacy laws and protect consumer's digital identity. The username/password security token used by service providers to authenticate, leaves the consumer vulnerable to phishing attacks.

- To solve above problems, a new technique emerged known as Identity-as–a-Service (IDaaS). IDaaS offers management of identity information as a digital entity. This identity can be used during electronic transactions.

(IAM)

Identity-as-a-Service may include the following:

- Form authentication
- Directory services
- Single sign-on services
- Federated services
- OpenID
- OAuth2
- SAML
- SCIM
- Provisioning & Deprovisioning

Cloud IAM typically includes following features:

- Single access control interface: a clean and consistent access control interface for all cloud services
- Enhanced security: One can define increased security for critical applications
- Resource-level access control: One can define roles and grant permissions to users to access resources at different granularity levels

Form authentication & Active Directory

Form authentication

- Register in a particular application to log in for authentication and authorization
- User have email-id and password, he/she enter user-id and password (in SSHA hashed format) into the login page
- The entered password is validated through database stored password
- Disadvantages :
  - Need to have account in every single organization
  - Can't interact to any third party services through LDAP

Active Directory

- A database that keeps track of all user accounts and password in your organization
- Directories tend to contain descriptive, attribute-based information
- Support filtering capabilities
- Allows you to store user accounts and passwords in location, improving your organizations' security

LDAP

1. Lightweight Directory Access Protocol (LDAP) is an Internet protocol used to access information directories.
2. A directory service is a distributed database application designed to manage the entries and attributes in a directory.
3. Runs over TCP/IP

LDAP vs Active Directory

- Ad is a Database
- Ldap is a protocol to access AD.

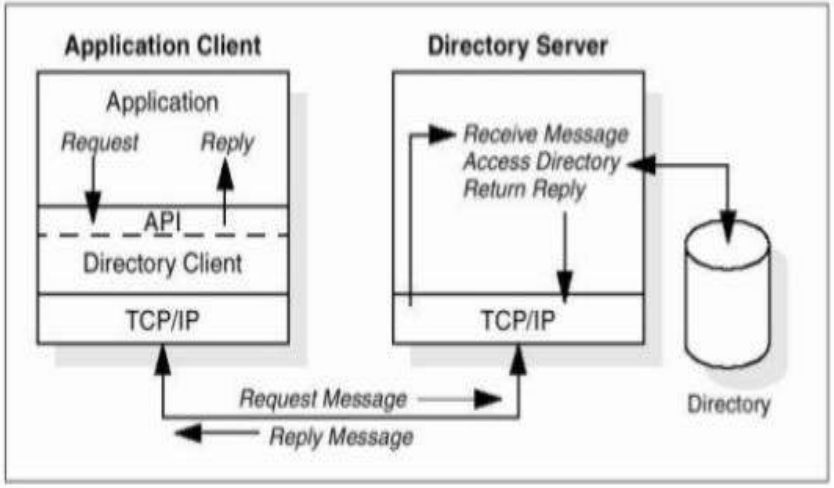
Eg. Microsoft Outlook work with directory services.

Directory structure

- A server holds a subtree starting from a specific entry, e.g. "dc=example,dc=com" and its children.
- Servers may also hold references to other servers
  - An attempt to access "ou=department,dc=example,dc=com" could return a *referral* or *continuation reference* to a server which holds that part of the directory tree.
- Client can then contact the other server
- Some servers also support *chaining*
  - Server contacts other server(s) and returns the results to the client

Directory Client/Server Interaction

- Clients performing protocol operations against servers
  - Client sends protocol request to server
  - Server performs operation on directory
  - Server returns response (results/errors)



Single sign-on

- Single sign-on - User Authentication Process
- One set of login credentials to access multiple applications
- Login into one application and access other apps without authenticating again

Single Sign-On/Off

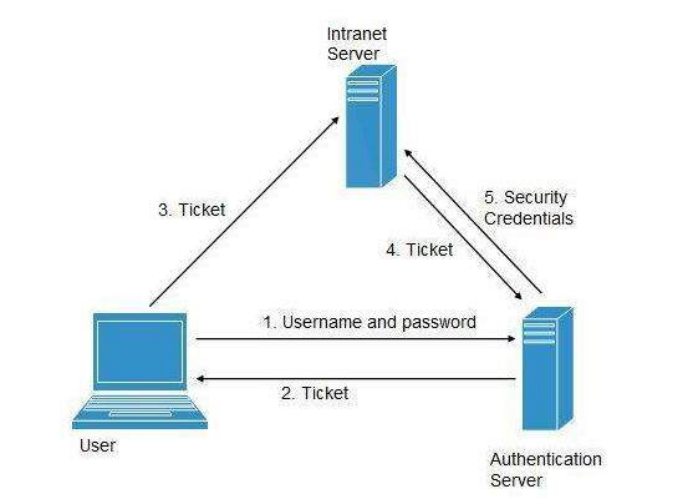
Integrated login through a Cloud company:

- Google
- Github
- Facebook
- Apple

Integrated corporate login:

- Okta
- Auth0
- VMware Workspace One Intelligent Hub
- UAA
- KeyCloak
- Any Openid Connect Certified provider

How does cloud computing work



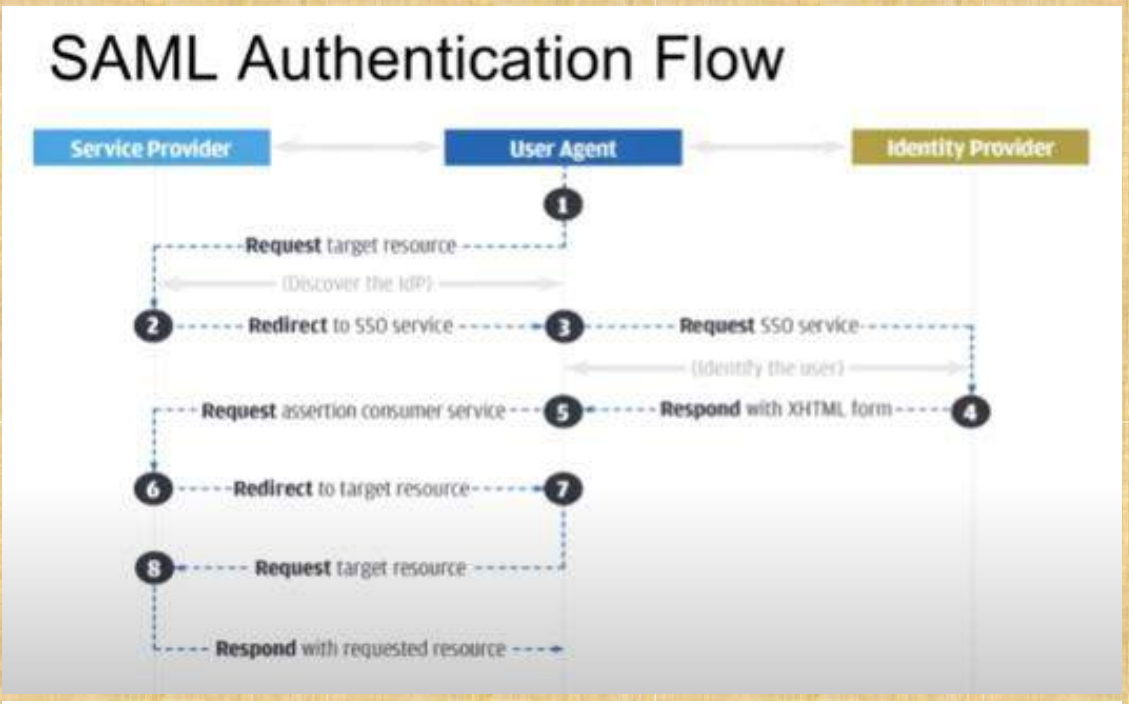
Advantages:

- User convenience : No need to remember too many passwords
- Less password fatigue caused by the stress of managing multiple passwords
- Less user time consumed by having to log in to individual systems
- Reduce help desk cost: Fewer calls to help desks for forgotten password
- Easy onboarding and offboarding

Disadvantages:

- Single source of failure

Security Assertion Markup Language (SAML)

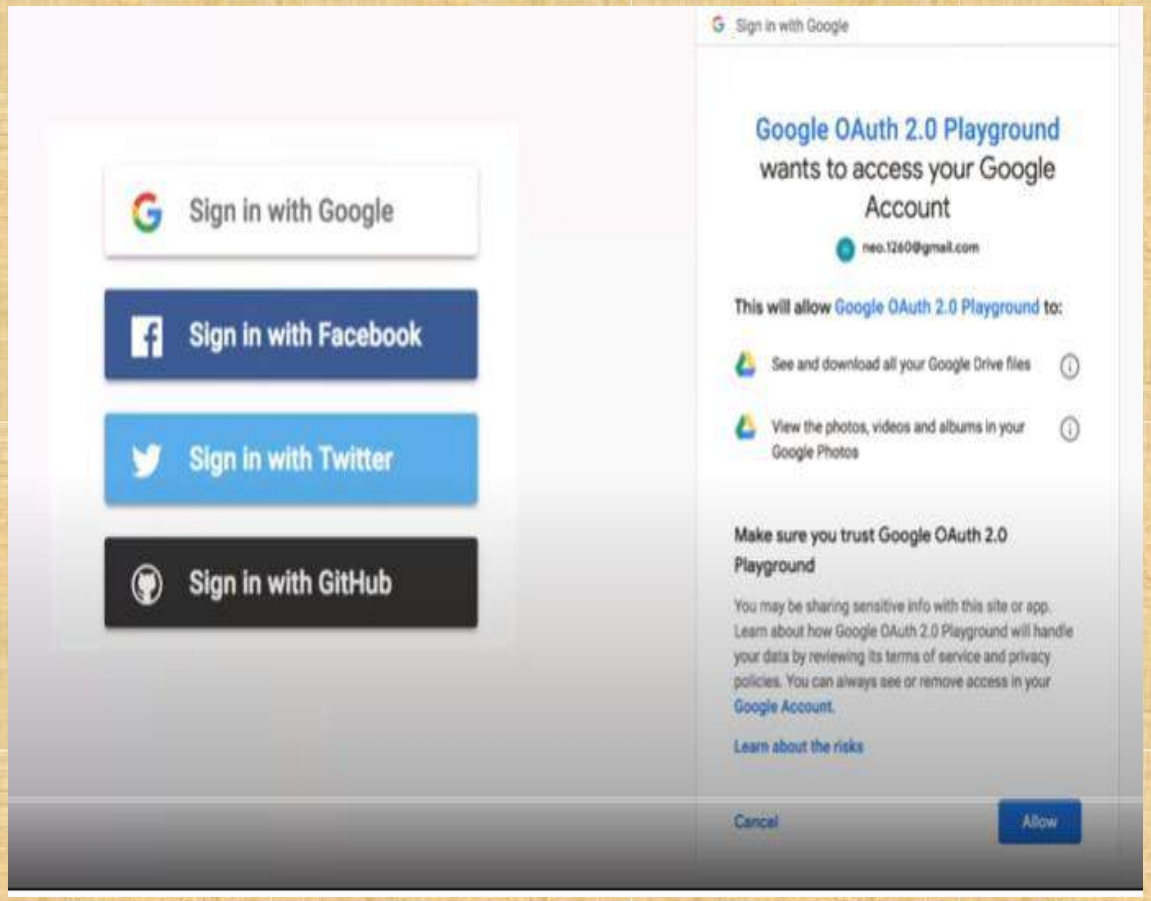


Few Observations

- Service Provider will never interact with Identity Provider
- Service Provider needs to know which Identity Provider it has to redirect the user to prior to authentication
- SAML authentication flow doesn't have to start from Service Provider. Identity Provider can also initiate it
- SAML authentication flow is asynchronous
- Service Provider does not maintain any state of any authentication requests generated

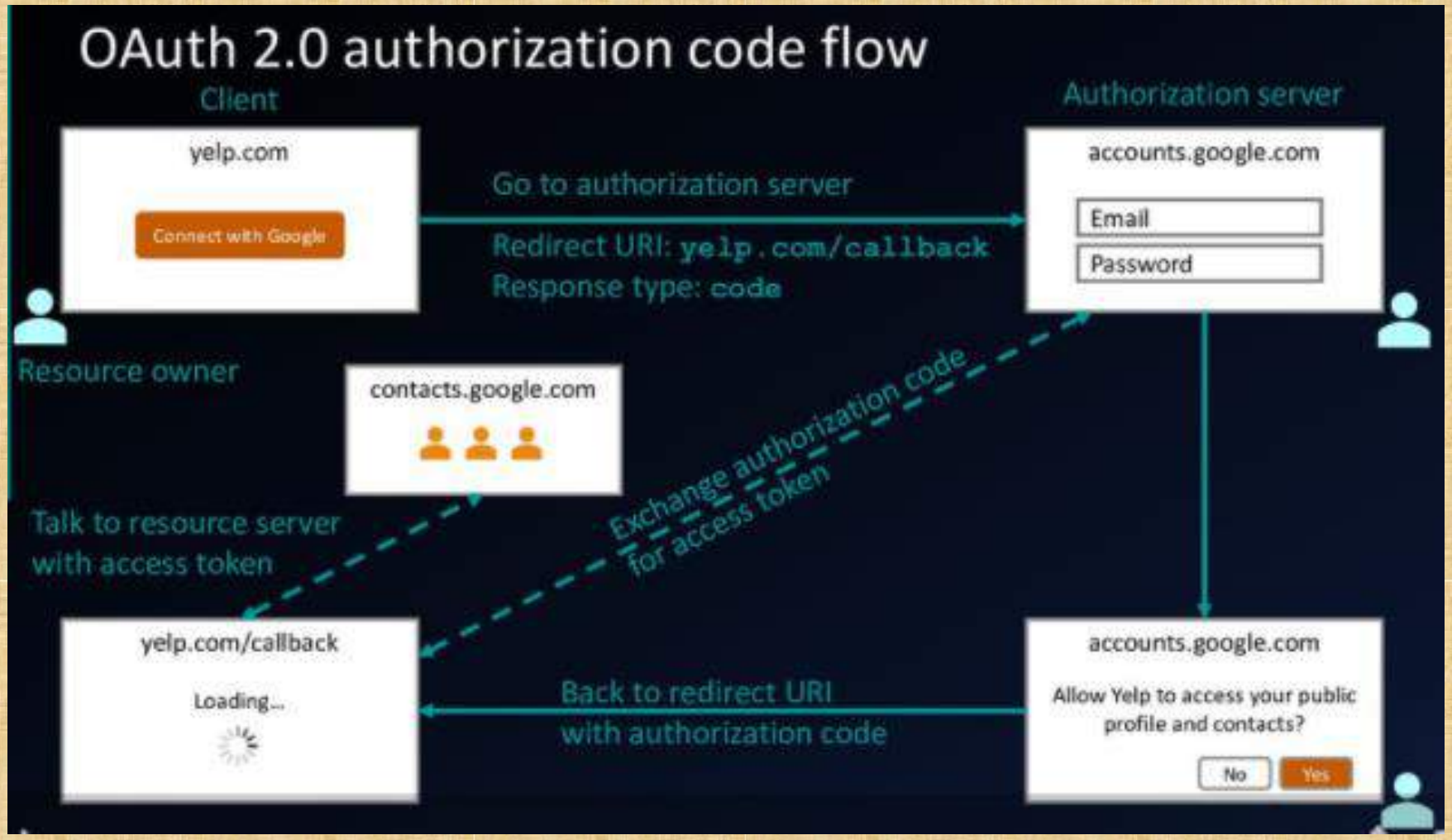
- XML based open standard protocol used by web browsers to enable Single Sign-On (SSO) through secure tokens.
- It does so by using standard cryptography and digital signatures to pass a secure sign-in token from an identity provider to an SaaS application.
- Most common SaaS vendors, such as Salesforce, Google and Microsoft already support SAML

OAuth 2.0 Protocol



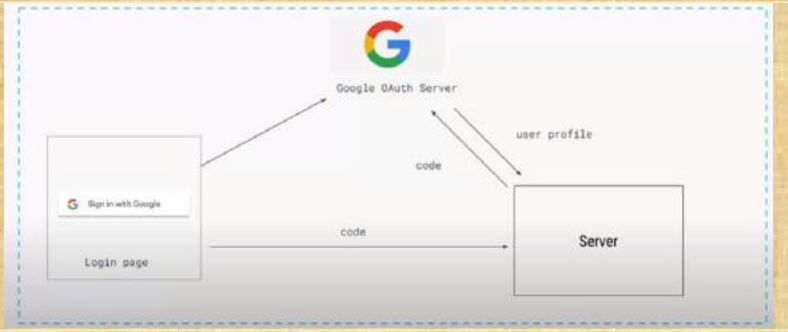
- A delegation or authorization protocol to provide (application/server) temporary access to a resource
- Works by delegating user authentication to the service that hosts the user account
- Provides authorization flows for web and desktop application, and mobile devices

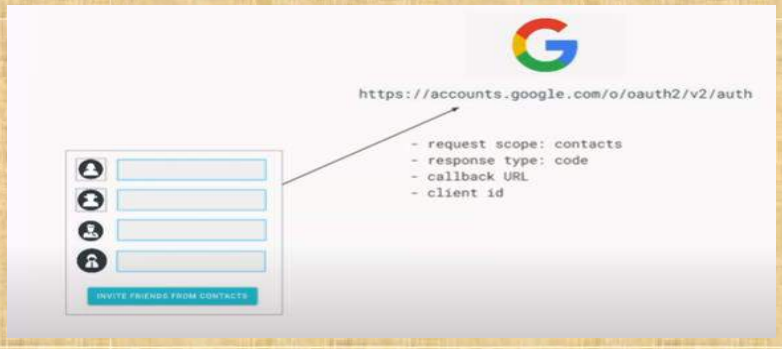
OAuth 2.0 Protocol

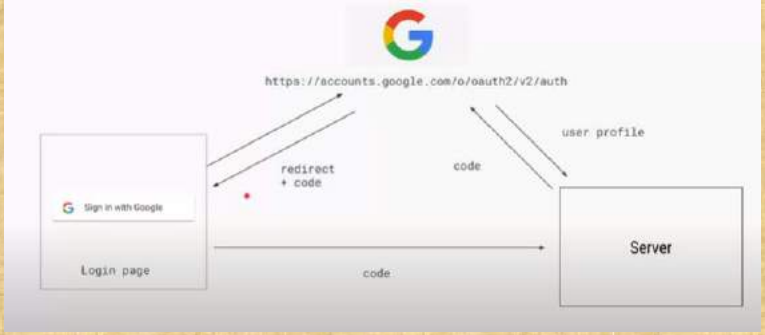


OpenID

- An open standard and decentralized protocol that support FDIM
- Allows end users to be authenticated by relying party using a third party provider
- A user can choose an Open ID identity provider (Google, Yahoo!, Flickr, MySpace, WordPress.com)
- Users' credentials are never provided to a relying party
- A relying party does not have to implement its own ad hoc login system
- It is built on Oauth2.0 protocol & uses additional JSON web token, called an ID token, to standardize area







Federated Identity Management (FDIM)

- FIDM describes the technologies and protocols that enable a user to package security credentials across security domains
- It uses Security Assertion Markup Language (SAML) to package a user's security credentials