# Unit IV: Data Acquiring , Organizing , Processing & Analytics

By: Er. Ishwar Rathod.

SCSIT, SUAS, Indore.

# Recall from previous units

- IoT / M2M architectural layers & functions learnt till are- devices communicate, first over a local network or WPAN and then send the physical layer data to data adaption & gateway layer.
- The gateway connects to the Internet, & communicates the data packets. The packets communicate over the Internet through a set of routers.
- Application & application-support layers use the acquired & collected data for IoT applications. Applications can also control & monitor the functions of devices. The application messages, commands & data communicates to the devices through the gateway using the Internet.

# Introduction

- Let us understand the functions required for applications, services & business processes at application-support & application layers.
- These functions are data acquiring, data storage, data transactions, analytics, results visualizations,
- IoT applications, services, processes, intelligence, knowledge discovery & knowledge management.

# Terms & their meaning used in IoT application layers.

- **Application** refers to application software or a collection of a software components. An application enables a user to perform a group of coordinated activities, functions & tasks.
- Streetlights control & monitoring is an **example** of an application. Software for tracking & inventory control are **other examples** of applications. Tracking applications use tags and locations data of the RFIDs.
- An application **enables** a user to withdraw cash using an Automatic Teller Machine (ATM). An umbrella sending warning messages for weather, A waste container management, health monitoring, traffic lights control, Synchronization and monitoring are other examples of IoT applications.

# Terms & their meaning used in IoT application layers.

- **Service** denotes a mechanism, which enables the provisioning of access one or more capabilities. An interface for the service provides the access to capabilities. The access to each capability is consistent with constraints & policies, which a service-description specifies.
- **Examples** of service capabilities are automotive maintenance service capabilities or service capabilities for the Automatic Chocolate Vending Machines (ACVMS) for timely filling of chocolates into the machines.
- Service consists of a set of related software components & their functionalities. The set is reused for one or more purposes. Usage of the set is consistent with the controls, constraints & policies which are specified in the service description for service. A service also associates a Service Level Agreement (SLA).
- A service consists of a collection of self-contained, distinct & reusable components. It provides logically grouped & encapsulated functionalities. Traffic lights synchronizing service, automobile maintenance service, device location, detection & tracking service, home security-breach detection & management service, waste containers substitution service, health-alerts service are the examples of IoT services.

# Terms & their meaning used in IoT application layers.

- **Service Oriented Architecture(SOA)** is a software architecture model, which consists of services, messages, operations & processes. SOA components are distributed over a network or the internet in a high level business entity. New business applications & application integration architecture in an enterprise can be developed using an SOA.
- **Message** means a communicating entity or objects.
- **Operation** means action or set of actions. For example, action during a bank transaction.
- **Transaction (trans + action)** refers to two interrelated set of operations or actions or instructions. For example a transaction may be access to sales data to select an get the annual sales in a specific year in return. One operation is access to sales data & other is annual sale in return. Another example of a transaction is a query transaction with a Data Base Management System.

# Terms & their meaning used in IoT application layers.

- **Query** is a command for getting select values from a data base which is return transfer the answer to the query after its processing. A query example is command to ACVMs data base for providing sales data of ACVMs on Sundays near city gardens in a specific festival period in a year. Another example is query to service centre data base for providing the list of automobile components needing replacement that have completed expected service life in specific vehicle.
- **Query processing** is a group of structure activities under taken to get the results from a data store as per the query.
- **Key Value Pair(KVP**) refers to a set of two linked entities, one is the key, which is a unique identifier for a linked entity & the other is the values, which is either the entity that is identified or a pointer in a location of that entity. **A KVP** example birthday – date pair. KVP is birthday: July 17, 2000. Birthday is the key for a table & date July 17, 2000 is the value. KVP applications create the look-up table, hash table & the network or device configuration files.

# Terms & their meaning used in IoT application layers.

- **Hash table(also called hash map)** refers to a data structure which maps the KVPs & is used to implement an associative array (for example array of KVPs). A hash table may use an index (key) which is computed using a hash functions & key map to the value. Index is used to get or point to the desired value.
- **Big table** maps two arbitrary string values into a associated arbitrary byte array. One is used as a row key & the other as column key. Time stamp associate in three dimensional mapping. Mapping is unlike a relational data base but can be considered as a sparse, distributed multidimensional sorted map. The table can scale up to 100s to 1000s of distributed computing nodes with ease of adding more nodes.
- **Business transaction (BT)** in data base theory, refers to a(business) process that request information from or that changes the data in a database. One operation in a BT is a command ‘ connect ‘ that connect a DBMS & database, which in turn also connects with the DBMS. Similarly, BTs are processes using commands ‘insert’, ‘delete’, ‘append’, & ‘modify’.
- **Process** means a composition of a group of structured activities or tasks that lead to a particular goal (or that interact to achieve a result). For example streetlight control process of the purchase process for a airline tickets. A process specifies activities with relevance rules based on data in the process.

# Terms & their meaning used in IoT application layers.

- **Process Matrix** is a multi-element entity, each element of which relates a set of data or inputs to an activity (subsets of activity).
- **Business process(BP)**is an activity or series of activity or a collection of interrelated structured activity, tasks or processes. A BP serves a particular goal  or specific results or service or product. The BP is a representation or process matrix or flowchart of sequence of activities with interleaving decision point; interleaving means putting in between. Decision point means an instance in a series of activities when decision are taken for further activities.
- **Business Intelligent (BI)** is a process which enables a business service to extract new fact & knowledge & then under take better decision. These new facts & knowledge follow from earlier results of data processing, aggregation & analysis of these results.

# DATA ACQUIRING AND STORAGE : **Data Generation**

- Data generates at devices that later on transfers to the Internet through gateway.
- **Passive device data** does not have its own power source. An external source helps such a device to generate and send data. Examples are RFID or ATM debit card.
- **Active device data** has its own power source Examples are Active RFID, streetlight sensor or wireless sensor node. Also has an associated microcontroller, memory & transceiver.
- **Event data** A device can generate data on an event only once. For example, on detection of traffic or on dark ambient conditions, which signals the event.
- **Device real-time data** An ATM generates data & communicate to the server instantaneously through the Internet. This initiates & enables Online Transactions Processing (OLTP) in real time.
- **Event –driven device data** A device can generate data on an event only once. Example a device receive command from controller or monitor, & then performs action(s) using an actuator. When action completes, then the device sends an acknowledgement.

# Data Acquisition

- Data acquisition means acquiring data from IoT /M2M devices. The data communicates after the interactions with a data acquisition system (application).
- The application interacts & communicate with a number of devices for acquiring the needed data.
- Application can configure sending of data after filtering or enriching at the gateway at the data adaption layer.
- Device –management software provisions for device ID or address, activation, configuring, registering, deregistering, attaching, & deattaching.

# Data validation

- Data acquired from the devices does not means that data are correct, meaningful or consistent.
- Data validation software do the validation checks on the acquired data.
- Validation software applies logic, rules & semantic annotations.
- Validation software consumes significant resources.
- An appropriate strategy need to be adopted. For example the adopted strategy may be filtering out the invalid data at the gateway or at device itself or controlling the frequency of acquiring or cyclically scheduling the set of devices in a industrial system. Data enriches, aggregates, fuses or compact at the adaptation.

# Data categorization for storage and Assembly software for the event

- Services, business processes & business intelligent use data. Valid, useful & relevant data can be categorized into three categories for storage--- data alone, data as well as result of processing, only the results of data analytics are stored.
- Logic 1 refers to an event generated but not acted upon. Logic 0 refers to an event generated and acted upon or not yet generated.

# Data store & Its features

- A data store is a data repository of a set of objects which integrate into the store.
- **Features-**
- Objects in a data-store are modeled using classes which are defined by the database schemas.
- A data stores is a general concept. It includes data repositories such as data base, relational database, flat file, spreadsheet, mail server, web server, directory services & VMware.
- A data store may be distributed over multiple nodes. Apache Cassandra  is an example of distributed data store.
- A data store may consist of multiple schemas or may consists of data in only one scheme. Example of only one scheme data store is a relational database.

# Data Centre Management

- Data centre is meant for data storage, data security and protection.
- A data centre is a facility which has multiple banks of computers, large memory systems, high speed network & Internet connectivity.
- The centre provides data security & protection using advanced tools, full data backups along with data recovery, redundant data communication connections & full system power as well as electricity supply backups.
- The manager of data centre is responsible for all technical & IT issues, operations of computers & servers, data entries, data security, data quality control, network quality control & the management of the services & applications used for data processing.

# Server Management

- Server management means managing services, setup & maintenance of system of all types associated with the server. A server needs to serve around the clock. Server management includes managing the following:
- Short reaction times when the system or network down.
- High security standards by routinely performing system maintenance & updation.
- Periodic system update for state of the art setup.
- Optimised performance.
- Monitoring of all critical service, with SMS & email notifications.
- Security of system & protection.
- Maintaining confidentiality & privacy of data.
- High degree of security & integrity & effective protection of data, files & databases at the organization.
- Protection of customer data or enterprise internal documents by attackers which includes spam mails, unauthorized use of the access to the server, viruses, malwares & worms.
- Strict documentation & audit of all activities.

# Spatial Storage

- Spatial Storage is storage as a Spatial data base which is optimized to store & later on receives queries from the applications.
- Internet communication by RFIDs, ATMs, Vehicles, Ambulances, Traffic light, Streetlights, Waste containers are example of where spatial database are used.
- A spatial data base can perform typical SQL queries, such as select statements & performs a wide variety of spatial operations. Spatial data base has the following **features:**

1. Can perform geometry constructor. For example creating new geometry.
2. Can define a shape using the vertices ( point or nodes).
3. Can perform observer function using queries which replies specific spatial information such as location of the center of a geometric object.
4. Can perform spatial measurement, which mean computing distance between geometries, length of line, areas of polygon & other parameters.
5. Can change the exiting feature to new ones using spatial functions & can predicate spatial relationship between geometries using true or false type queries.

# Organizing the data

- **Databases-**Required data values are organised as database(s) so that select values can be retrieved later.

- **Database-** one popular method of organising data is a database, which is collection of data. This collection organised into tables. A table provides a systematic way for access, management & update. A single table file is called flat file database. Each record is listed in separate row, unrelated to each other.

- **Rational Database-** A rational Database is a collection of data into multiple tables which relate to each other through special fields, called keys (primary key, foreign key & unique key). Relational databases provide flexibility. Examples of relational database are MySQL, PostGreSQL, Oracle database created using PL/SQL & Microsoft SQL server using T-SQL.
- Object Oriented Database (OODB) is a collection of objects, which save the objects in object oriented design. Examples are ConceptBase or Cache.

# Database Management System

- Database Management System (DBMS) is a software system, which contains a set of programs specially designed for creation & management of data stored in a database. Database transactions can be performed on a database or relational database.

# Atomicity, Data consistency, Data isolation & Durability (ACID) Rules

- The database transactions must maintain the atomicity, data isolation & durability during transactions.
- **Atomicity** means a transaction must complete in full, treating it as indivisible. When a service request completes, then the pending request field should also be made zero.
- **Consistency** means that data after the transactions should remain consistent. For example sum of chocolates that should equal the sums of sold & unsold chocolates for each flavour after the transactions on the database.
- **Isolation** means transactions between tables are isolated from each other.
- **Durability** means after completion of transactions, the previous transaction cannot be recalled. Only a new transaction can affect any change.

# Distributed Database

- **Distributed database (DDB)** is a collection of logically interrelated databases over a computer network. Distributed DBMS means a software system that manages a distributed database.
- The **features** of a distributed database system are:
- DDB is a collection of databases which are logically related to each other.
- Cooperation exits between the databases in a transparent manner. Transparent means that each user within the system may access all of the data within all of the databases as if they were a single database.
- DDB should be ‘location independent ‘, which means the user is unaware of where the data is located, it is possible to move the data from one physical location to another without affecting the user.

# Consistency, Availability & Partition-Tolerance Theorem

- **Consistency, Availability & Partition-Tolerance Theorem (CAP theorem)** is a theorem for distributed computing systems. The theorem states that it is impossible for a distributed computer system to simultaneously provide all three of the consistency, Availability, Partition tolerance (CAP) guarantees. This is due to the fact that a network failure can occur during communication among the distributed computing nodes. Partitioning of a network therefore needs to be tolerated. Hence, at all times either there will be consistency or availability.
- **Consistency** means 'Every read receives the most recent write or an error'. When a message or data is sought the network generally issues notification of time-out or read error. During an interval of a network failure, the notification may not reach the requesting node(s).
- **Availability** means **'Every** request receives a response, without guarantee that it contains the most recent version of the information'. Due to the interval of network failure, it may happen that most recent version of message or data requested may not be available.
- **Partition tolerance** means 'The system continues to operate despite an arbitrary number of messages being dropped by the network between the nodes'. During the interval of a network failure, the network will have two separate set of network nodes. Since failure can always occur therefore, the partitioning needs to be tolerated.

# Query Processing

- **Query** means an application seeking a specific data set from a data set from a database.
- **For example,** a query at a relational database at bank server may be for the ATM transactions made in a month by specific customer ID. **Other examples** are: most-liked chocolate flavour in the city by children of age group 6 to 10; number of times a vehicle visited at the ACPAMS center & **service** was rendered with satisfaction level of 5 out of 5.
- **Query Processing** means using a process & getting the results of the query made from a database. The process should use a correct as well as efficient execution strategy. Five steps in processing are:

1. **Parsing and translation:** This step translates the query into an internal form, into a relational algebraic expression & then parser, which check the syntax & verifies the relations.
2. **Decomposition** to complete the query process into micro-operations using the analysis (for the number of micro-operations required for the operations), conjunctive & disjunctive normalisation & semantic analysis.
3. **Optimisation** which means optimising the cost of processing. The cost means number of micro-operations generated in processing which is evaluated by calculating the costs of the sets of equivalent expressions.

## Query Processing & Distributed Query Processing

4. **Evaluation plan:** A query–execution engine (software) takes a query –evaluation plan & executes that plan.

5. **Returning** the results of query.

- **Distributed Query Processing** means query processing operations in distributed databases on the same system or networked systems. The distributed database system has the ability to access remote sites & transmit the queries to other systems.

# SQL

- **SQL** stands for Structured Query Language. It is a language for viewing or changing (update, insert or append or delete)databases. It is a language for data querying, updating, inserting, appending & deleting the databases. It is a language for data access control schema creation & modifications. It is also a language for managing the RDBMS.
- SQL was originally based upon the tuple relational calculus & relational algebra. SQL can embed within other language using SQL modules, libraries & pre-compilers.
- **SQL features are as follows**

▪ **Create schema** is a structure that contains description of objects created by a user (base tables, views, constraints). The user can describe & define the data for a database.

▪ **Create Catalog** consists of a set of schemas that constitute the description of the database.

▪ **Use Data Definition Language (DDL)** for the commands that depict a database, including, creating, altering & dropping tables & establishing constraints. The user can create & drop databases & tables

# NOSQL

- NOSQL stands for No-SQL or Not only SQL that does not integrate with applications that are based on SQL. NOSQL is used in cloud data store.

NOSQL may consists of the following:

- ❑ A class of non-relational data storage systems, flexible data models & multiple schema
- ❑ Class consisting of un-interpreted key & value of 'the big hash table '. For example in [Dynamo (Amazon S3)]
- ❑ Class consisting of unordered keys & using the JSON. For example in PNUTS
- ❑ Class consisting of ordered keys & semi-structured data storage systems. For example in the BigTable, Hbase & Cassandra (used in Facebook & Apache)
- ❑ Class consist of JSON. For example in MongoDb[6] which is widely used for NOSQL
- ❑ Class consisting of name & value in the text. For example in Couch DB
- ❑ May not require a fixed table schema

NOSQL system do not use the concept of joins (in distributed data storage systems). Data written at one node replicates to multiple nodes, therefore identical & distributed system can be fault-tolerant, & can have partitioning tolerance. CAP THEOREM is applicable. The system offers relaxation in one or more of the ACID & CAP properties. Out of the three properties (consistency, availability & partitions), two are at least present for an application.

- ❖ Consistency means all copies have same value like in traditional DBs.
- ❖ Availability means at least one copy available in case a partition becomes inactive or fails. For example, in web applications, the other copy in other partition is available.
- ❖ Partition means parts which are active but may not cooperate as in distributed databases.

# Extract, Transform & load

- Extract, Transform & load or ETL is a system which enables the usage of databases used, specially the ones stored at a data warehouse.
- **Extract** means obtaining data from homogeneous or heterogeneous data sources.
- **Transform** transforming & storing the data in an appropriate structure or format.
- **Load** means the structured data load in the final target database or data store or data warehouse.
- **All** the three phases can execute in parallel.
- **ETL SYSTEM** usages are for integrating data from multiple applications (systems) hosted separately.

# Relational Time Series Service

- Time series data means an array of numbers indexed with time(date-time or a range of date-time). Time series data can be considered as time stamped data. It means data carries along with it the date & time information about the data values. For example, sales of chocolates in Internet of ACVMs are different on different dates & times.
- Time series is any data-set that is accessed in sequence of time.
- Time Series Database (TSDB) is a software system which implements a database that optimally handles mathematical operations (profiles, traces, curves), queries or database transaction on time series.
- Conventional database systems, Relational Database System (RDMS) or flat file database software may not be modelled for time series handling.

# Real-Time & Intelligence

- Decision on real-time data is fast when query processing in live data (streaming) has low latency. Decision on historical data is fast when interactive query processing has low latency.
- Low latencies are obtain by various approaches. Massively Parallel Processing(MPP), in-memory databases & columnar databases.

# Transaction, Business Processes, Integration & Enterprise Systems

- **A transaction** is a collection of operations that form a single logical unit. For example, a database connect, insertion, append, deletion or modification transactions. Business transactions are transactions related in some way to business activity.
- **Online Transactions & Processing – OLTP** means process as soon as data or events generate in real time. OLTP is used when requirements are availability, speed, concurrency & recoverability in databases for real time data or events. Example on next slide gives the uses of OLTP in the application & network domain in Internet of ATMs ( ATM of a bank) connected to a bank server.

# Example

**Problem-**What are the usages of OLTP in the application & network domain in Internet of ATMs (ATM of a bank) connected to a bank server?

**Solution –**Server applications need processing & update-intensive database management with a high throughput. The requirements in these applications are availability, speed, concurrency & recoverability, & reduced paper trails. Therefore, the transactions at ATMs need OLTP.