EXCEL Homework 2

With the decreasing cost of providing technology and the increasing human activity on the internet comes an enormous increase in the amount of data collected. Researching a company you may like to interview with can often provide insight into the kinds of questions they might ask you in an interview. How does their business use data for its primary value? In other words why did they build their software?

# To empower you on your next interview here is a list of currently popular terms that describe how some of databases are being used today to fulfill their primary goal. Google them and collect a 1-3 sentence definition or description of each and include at least one use case for each. A use case is an example of how somebody will use it. A transactional database use case is: A person drives up to an ATM to withdraw cash from their bank account.

1. **Data warehouse**: a central repository for all or significant parts of the data that an enterprise's various business systems collect;

**Use Case**: real-time feeds and (often ad hoc) queries being made against the same set of data objects, daily or even hourly increases in data storage, purging of unneeded data,

1. **Transactional**: a database management system (DBMS) that has the capability to roll back or undo a database transaction or operation if it is not completed appropriately. It is defined for day to day operations like insert, delete and update. These databases are highly normalized (up to 3NF) to avoid redundancy.

**Use Case**: recording sales for a retail store.

1. **Analytical database**: also called an **analytical database**, is a read-only system that stores historical data on business metrics such as sales performance and inventory levels.

**Use Case:** Business analysts, corporate executives and other workers can run queries and reports against an **analytic database**.

1. **Time series database**: (TSDB) is a software system that is optimized for handling **time series** data, arrays of numbers indexed by **time** (a datetime or a datetime range). In some fields these **time series** are called profiles, curves, or traces.

**Use Case**: In the context of statistics, econometrics, quantitative finance, seismology, meteorology, and geophysics the primary goal of time series analysis is forecasting. In the context of signal processing, control engineering and communication engineering it is used for signal detection and estimation, while in the context of data mining, pattern recognition and machine learning time series analysis can be used for clustering, classification, query by content, anomaly detection as well as forecasting.

1. **Flat file database:** a database that stores data in a plain text file. Each line of the text file holds one record, with fields separated by delimiters, such as commas or tabs. While it uses a simple structure, a flat file database cannot contain multiple tables like a relational database can.

**Use Case:** WebDNA is a scripting language designed for the World Wide Web, with a hybrid flat file in-memory database system making it easy to build resilient database-driven websites

1. **Spatial database**: A spatial database is a database that is optimized to store and query data that represents objects defined in a geometric space. Most spatial databases allow representing simple geometric objects such as points, lines and polygons.

**Use Case:** Google Earth

1. Cloud database: Cloud platforms allow users to purchase virtual-machine instances for a limited time, and one can run a database on such virtual machines. With a database as a service model, application owners do not have to install and maintain the database themselves. Instead, the database service provider takes responsibility for installing and maintaining the database, and application owners are charged according to their usage of the service

**Use Case**: Amazon Web Services

# Here is a list of some popular database products being used today. Google them to discover how they distinguish themselves from the others. List at least one example of who is using it for what purpose, and one example of why it was chosen over the others.

[**Oracle**](https://en.wikipedia.org/wiki/Oracle_Corporation) **Exadata**

The main benefit is speed. Runs all types of databases, online transaction systems, processors and dwh, while solving poor performances of old database architecture.

**Who**

**Why**

We provide a complete solution to our customers from data center installation/configuration, firmware upgrades, ILOM upgrades, OS patching/upgrades, cluster patching/upgrades and database patching/upgrades. We upgrade the entire stack in a single evening with minimal outage. Depending on the customer’s tolerance for downtime, we can either perform the patching or upgrade in a rolling fashion.

Patching and upgrade services have proven to be a value-add differentiator for SMB and mid-market customers where resources and budgets are often limited. A single consulting resource can patch/upgrade the entire Exadata stack because patching and maintenance has become reliable and simplified.

A lot of our customers come from best of breed technologies (UCS/Dell/HP, EMC/Hitachi) to choosing Exadata. Typically, our customers choose Exadata for pure performance in IOPs, throughput, and low latency for their database workloads; however, we have seen a trend of customers choose the Exadata platform because they are short staffed, have high rates of attrition, and thus, have inability to support the hardware and software technologies.

[**MySQL**](https://en.wikipedia.org/wiki/MySQL)

Until its recent acquisition by Oracle, MySQL was possibly the most renowned open source database enterprise package that could be accessed completely for free. With the merger, the server increased both its features and its price tag, but there is still a free version available to the general community to contribute ideas and suggestions. MySQL is a fast and relatively inexpensive database management system. It can easily integrate with a wide variety of programing languages, and is considered a very reliable option. One of the most prominent features that customers seem to enjoy is the easy scalability of this system.

**Who**

Facebook, Tumblr, Scholastic, MTV Networks, Wikipedia, Verizon Wireless, Sage Group, Glassfish Open Message Queue, and RightNow Technologies

**Why**

MySQL provides a low cost and effective SQL Database solution for smaller companies that don’t require the overhead and cost associated with RDBMS tools like Oracle

[**Microsoft SQL Server**](https://en.wikipedia.org/wiki/Microsoft_SQL_Server)

SQL Server is the Microsoft-driven relational database management system. This system is used to store data as well as retrieve it when necessary; these functions can be supported by individual users or by multiple users within a larger network. The Microsoft SQL Server has warehousing options, quality and integration services, management tools that are simple to implement, as well as robust tools for development. Looking at the more technical end of things, Microsoft SQL Server uses query languages such as T-SQL and ANSI SQL. Disaster recovery is one of the product's most prominent features, in addition to in-memory performance, scalability, and corporate business intelligence capabilities.

**Who**

Microsoft SQL Server is used by businesses in every industry, including Great Western Bank, Aviva, the Volvo Car Corporation, BMW, Samsung, Principality Building Society, Wellmark Blue Cross and Blue Shield, and the Catholic District School Board of Eastern Ontario.

**Why**

It provides the ability to create an end-to-end automated BI platform through data management and provisioning, transformation and publication. Data platform development process improvement: Integrating the features of Visual Studio Team Foundation Server data projects for SQL Server development. Allows for smooth operation between development environments in conjunction with an agile process to release database changes

[**PostgreSQL**](https://en.wikipedia.org/wiki/PostgreSQL)

An object-relational database (ORDBMS) – i.e. a RDBMS, with additional (optional use) "object" features – emphasis on extensibility and standards compliance. As a database server, its primary function is to store data securely, and allow for retrieval at request of other software applications. Can handle workloads ranging from small single-machine to large Internet-facing applications (or for data warehousing) with many concurrent users; on macOS, PostgreSQL is the default database – for web hosting– and it is also available for Microsoft Windows and Linux (supplied in most distributions).

PostgreSQL is ACID-compliant and transactional, has updatable and materialized views, triggers, foreign keys; supports functions and stored procedures, and other expandability.

PostgreSQL is developed by the PostgreSQL Global Development Group, a diverse group of many companies and individual contributors. It is free and open-source software, released under the terms of the PostgreSQL License, a permissive free-software license.

**Who**

**Why**

PostgreSQL on the cloud is way cheaper than SQL Server on the cloud, and for more advanced queries, PostgreSQL is a much easier drop-in replacement for SQL Server than MySQL is.

doing spatial transformations with PostGIS (that SQL Server 2005 or 2008 can't do), and various other gyrations (not just for spatial but for financial apps and automated data categorization as well) that are made simpler with PostgreSQL array, advanced built-in text manipulation support (like regex,soundex, trigrams, levenshtein etc), and pushing this data back to a SQL Server database which is the database that drives much of the internal business processing.

With PostgreSQL you own your stack. If you use MySQL you need to be concerned about how you are deploying it and distributing or pay Oracle licensing ransom.

**IBM DB2**

contains [database server](https://en.wikipedia.org/wiki/Database_server) products developed by [IBM](https://en.wikipedia.org/wiki/IBM). These products all support the [relational model](https://en.wikipedia.org/wiki/Relational_model), but in recent years some products have been extended to support [object-relational](https://en.wikipedia.org/wiki/Object-relational) features and non-relational structures like JSON and XML. DB2 can be administered from either the command-line or a GUI. The command-line interface requires more knowledge of the product but can be more easily scripted and automated. The GUI is a multi-platform Java client that contains a variety of wizards suitable for novice users. DB2 supports both SQL and [XQuery](https://en.wikipedia.org/wiki/XQuery). DB2 has native implementation of XML data storage, where XML data is stored as XML (not as relational data or CLOB data) for faster access using XQuery.

**Who**

Many

**Why**

* Your company already maintains DB2 LUW or z/OS in house - this is a huge plus, as good DB2 skills are hard to find
* You have lots of storage available for z/OS mainframe systems - good SAN storage is not cheap and if you already have it available - use it
* Quality of optimizer - DB2's optimizer is industry's best and it is critical when it comes to optimizing performance of complex ad-hoc Data Warehouse queries
* Federation from multiple data sources - DB2 has superior data federation capabilities, but they come at extra cost
* Speed of data access - about 3x of Oracle and 6x of SQL Server  
  Better support comparing to other vendors
* Ease of administration and maintenance

**HP Vertica**

Most advanced SQL database analytics portfolio. Delivers speed w/out compromise, scale w/out limits, and the broadest range of consumption models. For on premise, on demand, in the cloud, or on Hadoop. Support for all leading BI and visualization tools, open source technologies like Hadoop and R, and built-in analytical functions, Derive more value from your Enterprise Data Warehouse, data lakes and analytics.

**Who**

BI Manager

**Why**

Fast columnar store database structure allows query times at least 10x faster than on any other database. Enables answers to data questions, numerous analytics on our data out to internal / external clients quickly. Integrate our Vertica data warehouse w/ Tableau to create reports quickly and efficiently. A two year backlog of report requests on old system virtually eliminated now that we are using Vertica to provide the solutions.

HPE Professional Services are well worth the investment.

Adding one node was very easy, as was adding memory to all nodes. We are currently in the process of setting up a Dev / DR environment which is going very smoothly.

**SAP Sybase ASE**

Transform your business and enhance in-the-moment decision making with the latest version of Sybase IQ Ð a highly optimized RDBMS built for extreme-scale Big Data analytics and warehousing. Gain an edge in todayÕs data-intensive world by providing deeper insights across your organization Ð insights never before practical, or even possible. Sybase IQ is a column based database, used for business intelligence, data warehousing, and data marts. It is optimized for BI/Analytics and ad-hoc queries. Since it is column based it uses optimal storage space and very high performance.

**Who**

ASR Group, Citrix, State of Indiana, PocketCard Co. Ltd

**Why**

Very few people are familiar with this technology, as training programs are very expensive.

We can now have over one billion rows in a table for semi-real trans-national querying. In quarterly or monthly report from billions of records, it supports and provides, stable performance.

Data compression ratio, mass query speed. High compression ratio and large query.

**Teradata**

A leading Massively Parallel Processing (MPP) architecture for data warehousing. Compare to IBM PureData. Teradata scales well but is also more expensive and requires more skilled DBA labor. The idea behind parallel processing is that “many hands make light work”. In other words large tasks become small when divided among several people. In this context “people” are processing units of memory, CPU and disk. MPP is when you get 64, 128, 256, or more synchronized processing units executing a database query simultaneously. The primary design philosophy of both Netezza and Teradata is simple. Each processor has total control and responsibility for a disk. Rows of every table are spread or distributed as evenly as possible across all the disks in the system and then data can be retrieved in parallel.

Teradata is a relational database management system (RDBMS) that is:  
• Open system, running on a UNIX MP-RAS or Windows server platform.  
• Capable of supporting many concurrent users from various client platforms.  
• Teradata is compatible with industry standards (ANSI compliant).  
• Teradata is completely built on a parallel architecture.

**Who**

**Why**

* There have plenty of reasons why customers like to choose Teradata
* Teradata supports more larger warehouse data than all competitors combined.
* Teradata Database can scale from 100 gigabytes to over 100+ petabytes of data on a single system without losing any performance .This is called Scalability.
* Provides a parallel-aware Optimizer that makes query tuning unnecessary
* Automatic and even data distribution eliminates complex indexing schemes or time-consuming reorganizations .
* Teradata Database can handle the most concurrent users, who are often running multiple, complex queries.
* Designed and built with parallelism.
* Supports ad-hoc queries using SQL
* Single point of control for the DBA (Teradata Manager).
* Unconditional parallelism (parallel architecture)
* Teradata provides the lowest total cost (TCO) of ownership
* High availability of data because there is no single point of failure - fault tolerance is built-in to the system.

Teradata Database can be used as

* Enterprise data warehousing
* Active data warehousing
* Customer relationship management
* Internet and E­Business
* Data marts

# The class focuses on learning SQL using Oracle. All of the databases above use SQL. The instructor postulates what you learn in this class is 95-97% transferrable to any of the products listed that in a typical job setting is. Is that true? Or, is the instructor full of beans and gravy.

# Founded in 1918, ANSI (American National Standards Institute) is a private, non-profit organization that administers and coordinates the U.S. voluntary standards and conformity assessment system. Periodically ANSI updates and republishes the “Database Languages – SQL” standards. Each manufacturer claims ANIS compliance with these standards.

# Below is a table showing syntax differences between Oracle SQL and Microsoft SQL Server SQL. Although you may intuit meaning from their names, we have not covered any of these and you are not expected to understand them. Three of the columns are blank. Fill in these three columns.

**SQL Functions: Description and Syntax**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Description** | **Oracle** | **MS SQL Server** | **PostgreSQL** | **MySQL** | **Vertica** |
| Find smallest integer >= n | CEIL | CEILING | CEILING | CEILING | CEILING |
| Modulus | MOD | % | MODULE | MOD | MOD |
| Truncate number | TRUNC | <none> | TRUNCATE | TRUNCATE | TRUNC |
| Translate NULL to n | NVL | ISNULL | ISNULL | IS\_NULL | ISNULL |
| Return NULL if two values are equal | DECODE | NULLIF | NULLIF | NULLIF | NULLIF |
| String concatenation | CONCAT(str1,str2) | str1 + str2 | string || string | CONCAT | CONCAT |
| Capitalize first letters of words | INITCAP | <none> | UPDATE lala\_table SET lala\_col=initcap(lower(lala\_col)) | UPPPER ? | INITCAP |
| Find string in string | INSTR | CHARINDEX | SELECT id FROM TAG\_TABLE WHERE 'aaaaaaaa' LIKE '%' || tag\_name || '%'; | INSTR | INSTR |
| Find pattern in string | INSTR | PATINDEX | ‘abc’ LIKE ‘abc’ true  ‘abc’ LIKE ‘a%’ true  ‘abc’ LIKE ‘\_b\_’ true  ‘abc’ LIKE ‘c’ false | **FROM**  **LIKE** | MATCH |
| String length | LENGTH | DATALENGTH | LENGTH | LENGTH | CHARACTER\_LENGTH |
| Pad string with blanks | RPAD, RPAD | <none> | PAD | RPAD | RPAD |
| Trim leading or trailing chars other than blanks | LTRIM, RTRIM, TRIM | <none> | TRIM | TRIM | TRIM, TTRIM, LTRIM |
| Replace chars in string | REPLACE | STUFF | REPLACE | REPLACE | REPLACE |
| Convert number to string | TO\_CHAR | STR, CAST | CAST | CAST | CAST |
| Convert string to number | TO\_NUMBER | CAST | CAST | CAST | CAST |
| Get substring from string | SUBSTR | SUBSTRING | SUBSTRING | SUBSTRING | SUBSTRING |
| Date addition | ADD\_MONTH or + | DATEADD | + | ADDDATE | TIMESTAMPADD |
| Date subtraction | MONTHS\_BETWEEN or - | DATEDIFF | - | DATEDIFF | TIMESTAMPDIFF |
| Last day of month | LAST\_DAY | <none> | LAST\_DAY | LAST\_DAY | LAST\_DAY |
| Time zone conversion | NEW\_TIME | <none> | NEW | CONVERT\_TZ | NEW\_TIME |
| Next specified weekday after date | NEXT\_DAY | <none> | NEXT |  | NEXT\_DAY |
| Convert date to string | TO\_CHAR | DATENAME, CONVERT | TO\_CHAR |  | TO\_CHAR |
| Convert string to date | TO\_DATE | CAST | CAST | STR\_TO\_DATE | TO\_DATE |
| Convert date to number | TO\_NUMBER(TO\_CHAR(d)) | DATEPART | To\_number | TO\_DAYS | TO\_NUMBER |
| Date round | ROUND | CONVERT | ROUND | ROUND | ROUNDDATE |
| Date truncate | TRUNC | CONVERT | TRUNCATE |  | DATE\_TRUNC |
| Current date | SYSDATE | GETDATE | CURRENT\_DATE | CURDATE | CURRENT\_DATE |
| If statement in an expression | CASE, DECODE, COALESCE | CASE, COALESCE | CASE, COALESCE | CASE | CASE |
| Current user | USER | USER | CURRENT\_USER | CURRENT\_USER | USERS |