Informatica PC Training

Day-1

Agenda:

- Introduction to Data Warehousing
- Informatica Architecture & Tool Suite
- PowerCenter Tool Stack

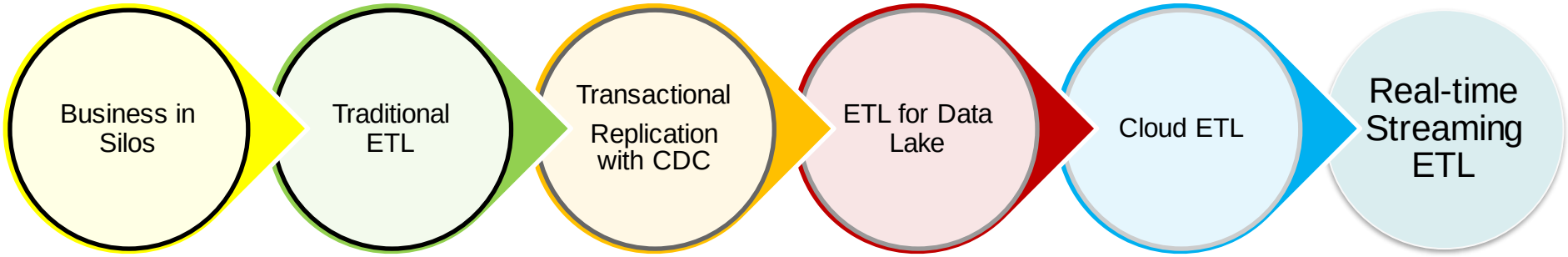
Debadatta Mohanty

HCL

Housekeeping Tips

- **Please mute your phone during the presentation.**
- **If there is too much noise, participants will be put on auto-mute.**
- **We shall open up the table for Q&A at the end of the session.**
- **Please feel free to post your questions over Chat as well.**
- **This session will be recorded and an email will be sent with links to the recordings after the session.**
- **At the end of the course, TEX will request you to provide feedback on the training.**

ETL Evolution



Business in Silos
- Dark ages- Pre 2k
- Silo mode Data Retrieval, Load, Backup

Traditional ETL
- Batch-based, higher latency
- On-Disk/On-Premise advanced transformations
- Designed mainly for Databases

CDC
- Real-time or low-latency
- Lack of transformation capabilities
- Designed mainly for Databases

Data Lake
- Batch or real-time
- On-Disk/On-Premise advance transformations
- Retrieve, Process, Load BigData
- Automatic Schema Detection

Cloud ETL
- Batch, CDC, Real-time
- Pay as you use
- Quick Setup
- Pre-build templates
- Retrieve, Process, Load from On-Premise to Cloud

Streaming ETL
- Combines low-latency of replication with transformation capabilities of ETL
- Performs in-line transformations in-memory
- Designed to support varied sources, targets

HCL

Introduction to Data Warehousing

What is Data Warehouse?

A relational database that is designed for query and analysis.

It usually contains historical data derived from transaction data, but it can include data from other sources.

It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources.

In addition to a relational database, a data warehouse environment includes an extraction, transportation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

A data warehouse is a collection of corporate information and data derived from operational systems and external data sources.

HCL

Introduction to Data Warehousing

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon:

- **Subject Oriented**
- **Integrated**
- **Nonvolatile**
- **Time Variant**

Subject Oriented:

DWH are designed to help you analyze data. E.g, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes it subject oriented.

Integrated:

Integration is closely related to subject orientation. DWH must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile:

Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

Time Variant:

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

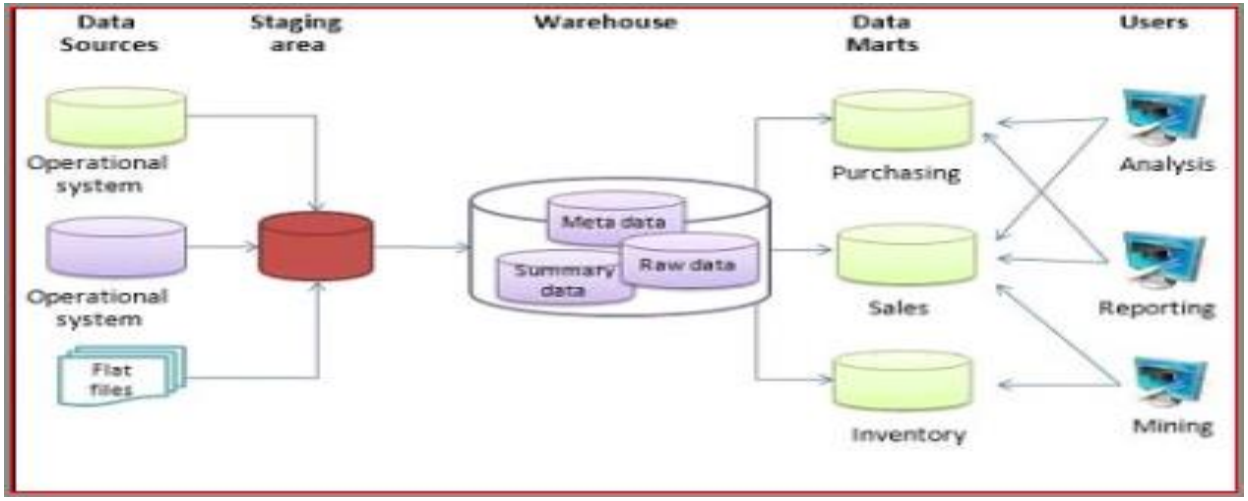
HCL

Data Warehouse Architecture

In general, all data warehouse systems have the following layers:

- Data Source Layer
- Staging Area
- ETL Layer
- Data Storage Layer
- Data Presentation Layer

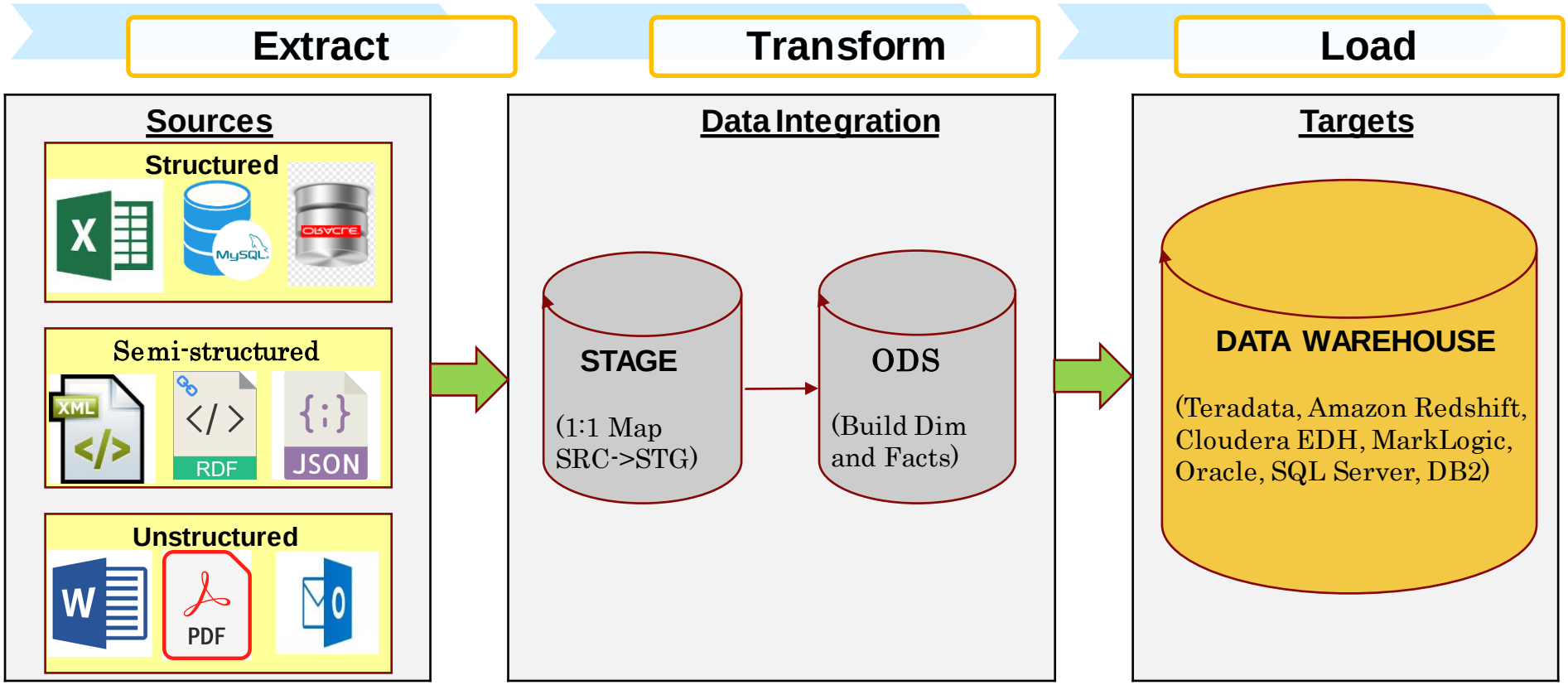
Basic Architecture of DWH



Datawarehouse Databases

- **Teradata:** It is market leader in data warehouse space. Teradata's EDW (enterprise data warehouse) platform provides businesses with robust, scalable hybrid-storage capabilities and analytics from mounds of unstructured and structured data leading to real-time business intelligence insights, trends, and opportunities.
- **Oracle:** Oracle 12c Database is the industry standard for high performance scalable, optimized data warehousing. The Oracle Exadata Database Machine is engineered to deliver dramatically better performance, cost effectiveness, and availability for Oracle databases. Exadata features a modern cloud-based architecture with scale-out high-performance database servers, scale-out intelligent storage servers with state-of-the-art PCI flash, and an ultra-fast InfiniBand internal fabric that connects all servers and storage.
- **Amazon Web Services (AWS):** The whole shift in data storage and data warehousing to the cloud over last several years has been momentous and Amazon is market leader in this space. E.g. Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse solution.
- **Cloudera:** It has emerged as a major enterprise provider of Hadoop-based data storage and processing solutions. Cloudera offers an Enterprise Data Hub (EDH) for its variety of operational data store or data warehouse. EDH is a framework for "information-driven enterprise" and focused on "batch processing", interactive SQL, enterprise search and advanced analytics, Cloudera's data warehouse is based on CDH, which is cloudera' s version of Hadoop.
- **MarkLogic:** It offers NoSQL database platform. This is an alternate forms of storage and processing. It released a new semantics platform which provides capability of storing billions of RDF triples that queried with SPARQL (a semantic query language for RDF platform) to provide richer, deeper insights to data in ways not possible within relational models.

ETL Process Flow



ETL Tools & Role of Informatica in DWH

Top 5 ETL tools in market

1) Informatica – PowerCenter.
2) IBM – Infosphere Information Server.
3) Ab Initio.
4) Microsoft – SQL Server Integrated Services (SSIS)
5) Talend

Informatica plays role in Extracting, Transforming and Loading data in DWH

Informatica Power Centre 10.2 Suite

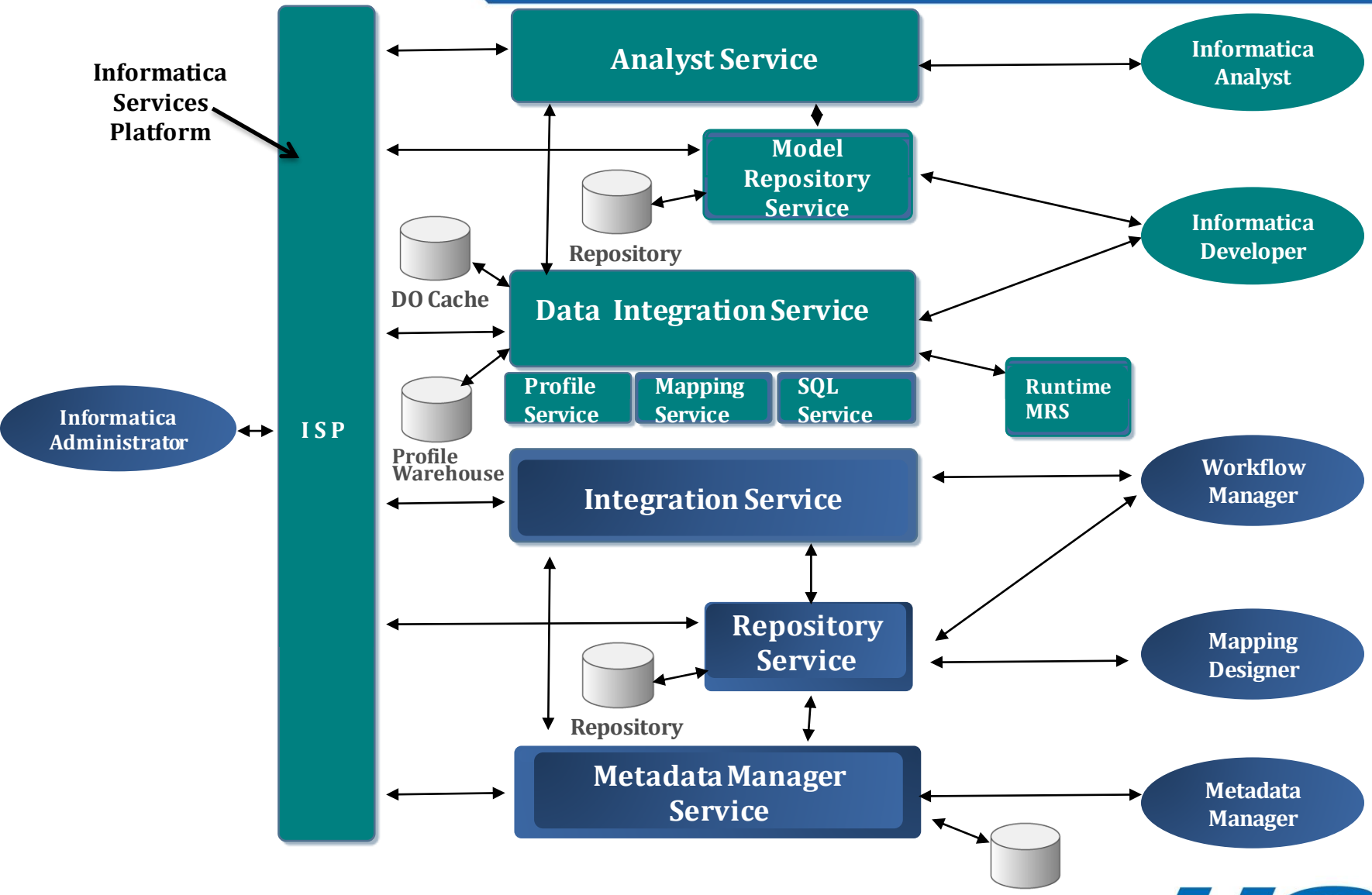
❑ Power Centre Server

- Repository Service
- Integration Service

❑ Power Centre Client

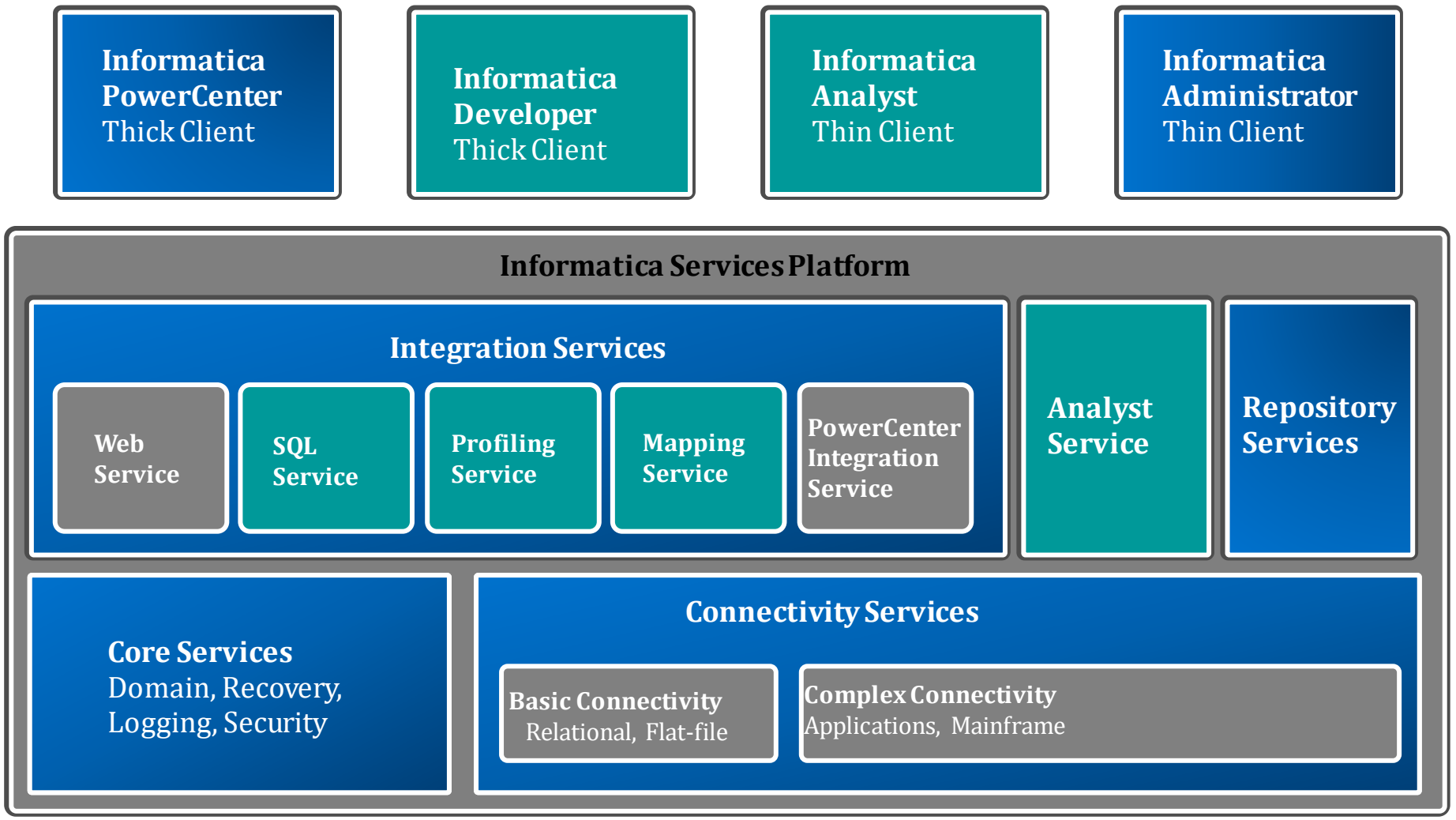
- Administration Console
- Power Centre Repository Manager
- Power Centre Designer
- Power Centre Workflow Manager
- Power Centre Workflow Monitor
- Power Centre Mapping Architect for Visio

Informatica 10.2 Architecture

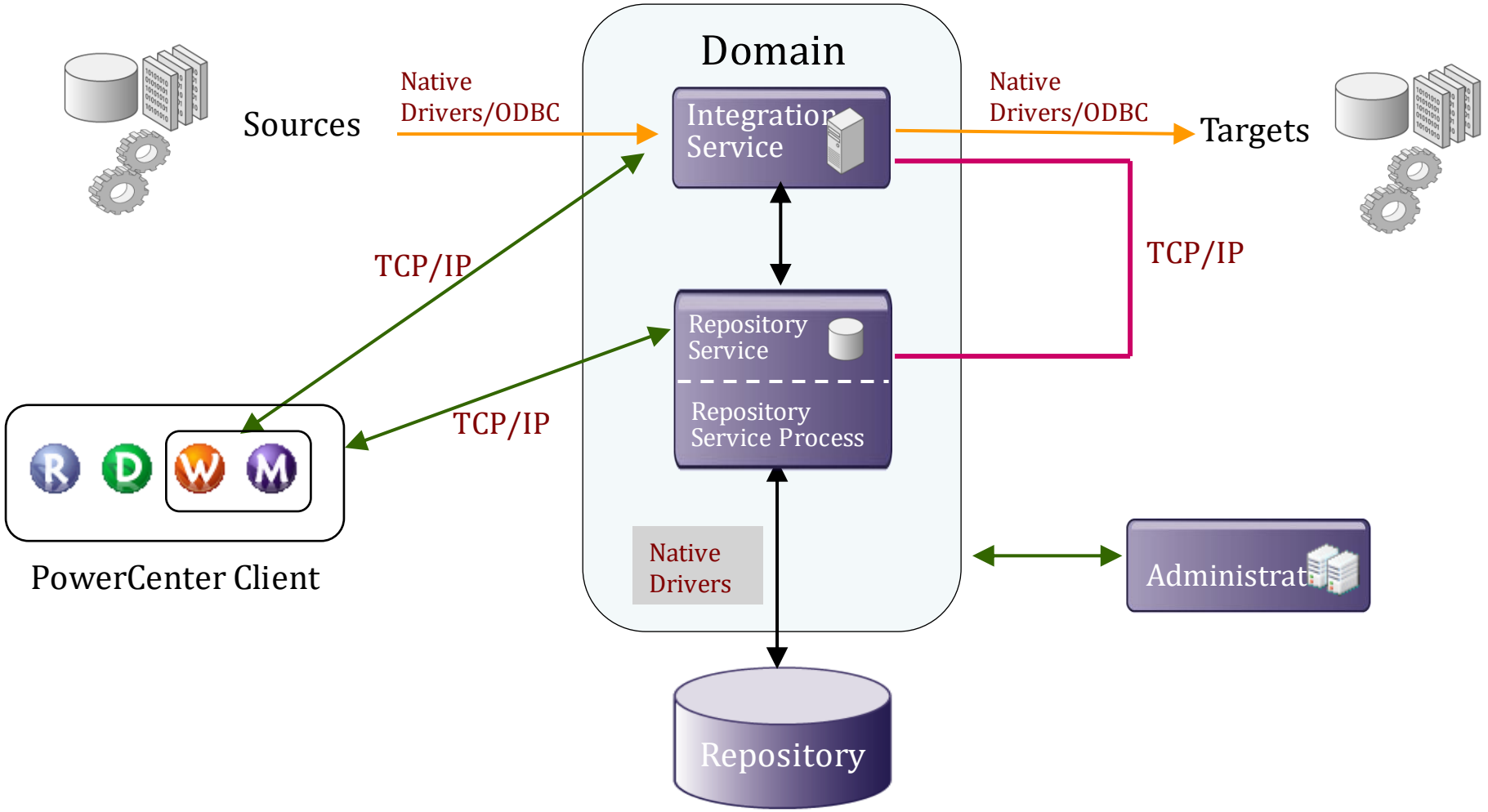


Informatica Platform Architecture

Evolving the Platform with Version 9.x

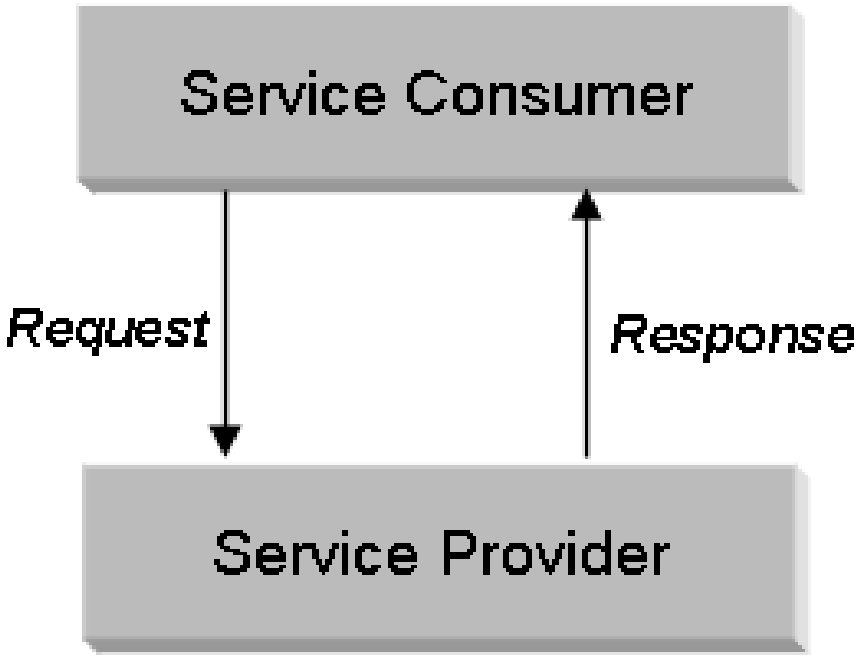


Informatica Connectivity



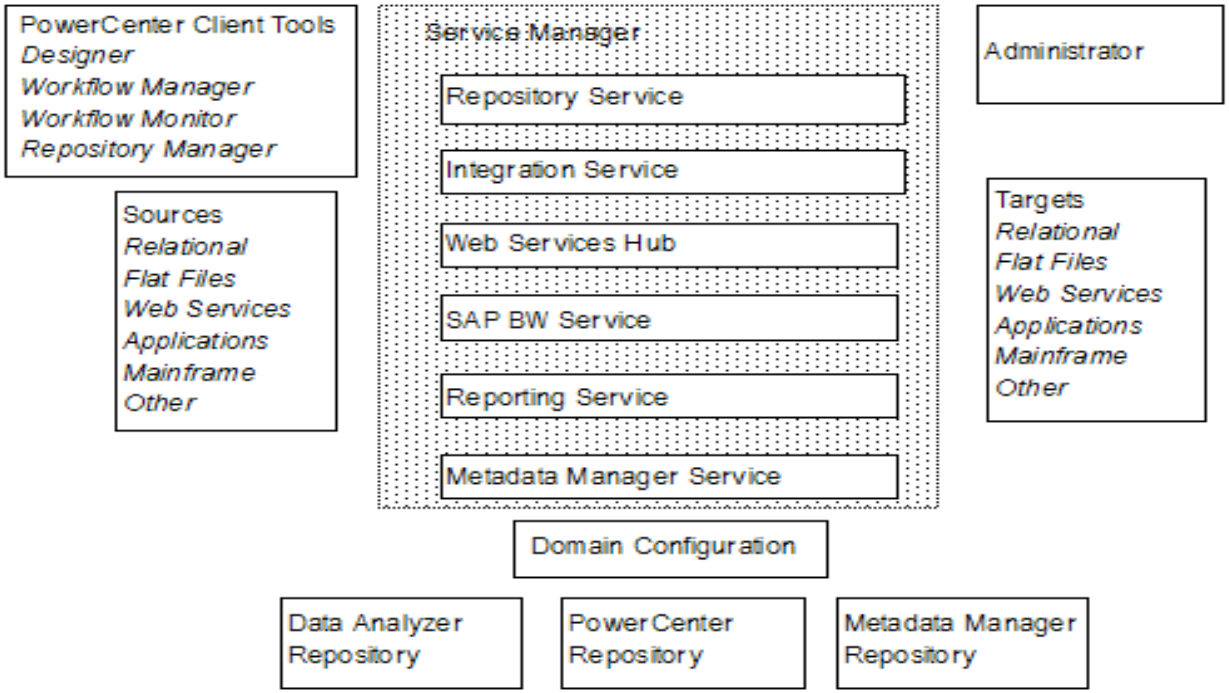
Service Oriented Architecture

- SOA: An application architecture in which all functions, or services, invoke software interfaces that perform business processes.
- Service: A task performed by a service provider to achieve desired end results for a service consumer. Both provider and consumer are roles played by software agents on behalf of their owners.



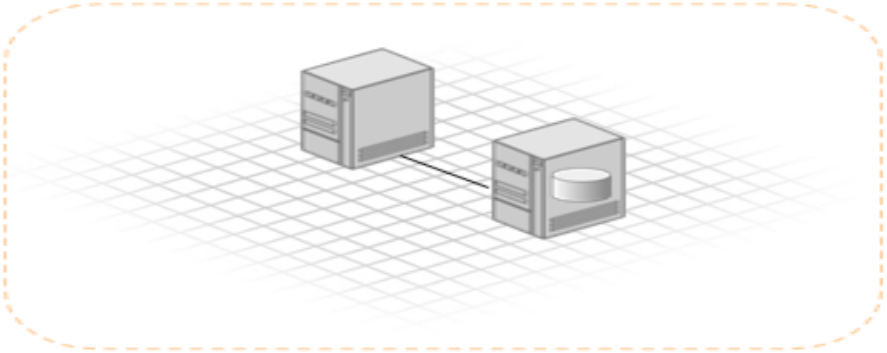
HCL

Informatica Power Centre Server components



Domain, Nodes and Services

- **Domain can be a single PowerCenter installation**
  - e.g. single Integration Service, Repository Service



- **A collection of nodes and services**
- **Primary unit of administration**

HCL

Nodes

- The Logical Representation of Machine in Domain.

- There could be “n” no of nodes in Informatica Domain.

- Node is configured to run the application services such as Repository service and Integration Service.

Service Manager

- Service Manager is built into Domain to Support the domain and application Services.

- Service Manager runs on each node in the Domain.

- SM Starts and runs the application Services on a Machine.

- SM performs Alerts, Authentication, Authorization , Domain configuration, Node configuration, Licensing , Logging and User Management.

Application Services

A group of services that represent Informatica server-based functionality. The application services that run on each node in the domain depend on the way you configure the node and the application service.

- Analyst Service.
- Data Integration Service.
- Model Repository Service.
- PowerCenter Repository Service.
- PowerCenter Integration Service.
- Web Services Hub.
- Reporting Service.
- Metadata Manager Service.

PowerCenter Services

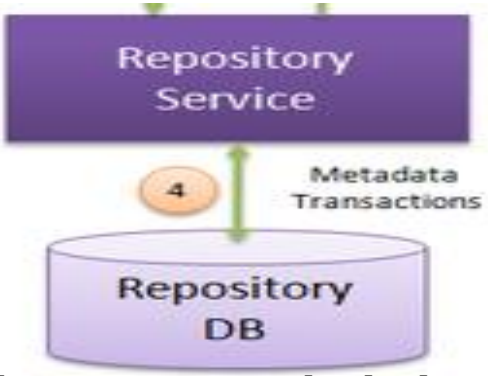
- **PowerCenter Repository Service**: The PowerCenter Repository Service manages connections to the PowerCenter repository from repository clients. A repository client is any PowerCenter component that connects to the repository The Repository Service is a separate, multi-threaded process that retrieves, inserts, and updates metadata in the repository database tables. The Repository Service ensures the consistency of metadata in the repository.

- **PowerCenter Integration Service:** The PowerCenter Integration Service reads workflow information from the repository. The Integration Service connects to the repository through the Repository Service to fetch metadata from the repository

- Both services should up and running to run workflow.

PowerCenter Repository

The PowerCenter repository resides in a **relational database**. The repository stores information required to extract, transform, and load data. It also stores administrative information such as permissions and privileges for users and groups that have access to the repository.



Global repository. The global repository is the hub of the repository domain. Use the global repository to store common objects that multiple developers can use through shortcuts. These objects may include operational or application source definitions, reusable transformations, mapplets, and mappings.

Local repositories. A local repository is any repository within the domain that is not the global repository. Use local repositories for development. From a local repository, you can create shortcuts to objects in shared folders in the global repository. These objects include source definitions, common dimensions and lookups, and enterprise standard transformations. You can also create copies of objects in non-shared folders.

PowerCenter Designer

- **Source Analyzer**. Import or create source definitions.

- **Target Designer.** Import or create target definitions.

- **Mapping Designer.** Create mappings that the Integration Service uses to extract, transform, and load data. You can display the following windows in the Designer

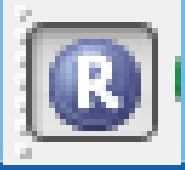
- **Transformation Developer.** Develop transformations to use in mappings. Reusable transformation can be used in multiple mappings.

- **Mapplet Designer.** Create sets of transformations to use in mappings.

- **Navigator:** Connect to repositories and open folders within the Navigator. You can also copy objects and create shortcuts within the Navigator.

- **Output.** View details about tasks you perform, such as saving your work or validating a mapping.

Repository Manager



- **Manage user and group permissions**. Assign and revoke folder and global object permissions.

- **Perform folder functions.** Create, edit, copy, and delete folders. Work you perform in the Designer and Workflow Manager is stored in folders. If you want to share metadata, you can configure a folder to be shared.

- **View metadata.** Analyze sources, targets, mappings, and shortcut dependencies, search by keyword, and view the properties of repository objects.

Workflow Manager

- **Workflow Designer**. Create a workflow by connecting tasks with links in the Workflow Designer. You can also create tasks in the Workflow Designer as you develop the workflow.

- **Worklet Designer.** Create a worklet in the Worklet Designer. A worklet is an object that groups a set of tasks. A worklet is similar to a workflow, but without scheduling information. You can nest worklets inside a workflow.

- **Task Developer:** Create tasks you want to accomplish in the workflow.

Workflow Monitor

- **Navigator window :** Displays monitored repositories, servers, and repositories objects.

- **Output window:** Displays messages from the Integration Service and Repository Service.

- **Time window:** Displays progress of workflow runs.

- **Gantt Chart view:** Displays details about workflow runs in chronological format.

- **Task view:** Displays details about workflow runs in a report format.

Demo and Q&A

Overview of

- **Informatica Admin console**
- **Informatica PowerCenter Client tools**

Thank You