kagool

Velocity: Big Data Ingestion & Management using Azure Cloud Services

For SAP and Non-SAP Landscapes

Version 1.1
27-Feb-2019

Document Control

Version

Version	**Date**	**Author(s)**	**Comments**
0.1	25-Sept-18	Kagool Team	Initial Version created and shared with MS for review
0.2	26-Sept-18	Kagool Team	Updated following comments from the team
1.0	26-Sept-18	Kagool Team	Baselined
1.1	27-Feb-19	Kagool Team	Updated with latest developments
1.2	27-May-19	Kagool Team	Updated with latest developments

Contributors

Name	**Company**	**Role**	**Contact**
Innovation Lab Team	Kagool	Technical Director	Prashant Patel (ppatel3@kagool.com)
Mahesh Balija	Microsoft	Cloud Solution Architect	mahesh.balija@microsoft.com
Pankaj Meshram	Microsoft	Cloud Solution Architect	pankaj.meshram@microsoft.com
Neil Gardner	Microsoft	SAP Alliances Manager	neil.gardner@microsoft.com
Chandu Cheeti	Microsoft	Azure Strategy Architect Lead	chandu.cheeti@microsoft.com

Approvals

Version	**Date**	**Context**	**Approved By**
1.0	26-Sept-18	Final	Kagool & MS team
1.1	tbc	tbc	tbc

Distribution

Group	**Distribution**
Kagool Team	Dan Barlow, Prashant Patel, Hitesh Bansal, Suresh Kanumuri, Akshata Ankam, Anuj Mutha, Shiva Gavini, Kalyan Gupta, Naresh Reddy, Rupa Kumari, Sameer Chitre, Sunitha Debbadi, Sarvadnya Mutalik, Sandra Owens, Umesh Yadav
Microsoft Team	Mahesh Balija, Chandu Cheeti, Pankaj Meshram, Neil Gardner

Introduction

Kagool is recognised as specialists in migrating and extracting large complex data sets into and out of SAP systems. Having worked in a number of Industry verticals, Kagool has created many bespoke solutions to deal with complex integration patterns.

Microsoft contacted Kagool for an open discussion around the challenges of moving large data volumes from SAP & non-SAP systems into the MS Azure Data Lake in an efficient manner as close to real time as possible

Kagool agreed to partner with Microsoft to create a high performing data syndication solution using SAP & Azure technologies that would support the following capabilities:

- Simple to deploy
- Simple to use
- Be reusable without the need for continuous developments
- Support Full CDC/Delta management
- Offer near-real time CDC

The Kagool Innovation Lab worked under the technical guidance & leadership of the MS Azure COE architects (Mahesh Balija, Pankaj Meshram, Neil Gardner, Chandu Cheeti), to develop a POC which has matured into a robust and scalable solution that is now deployed on major SAP client estates.

This data ingestion solution is called **Velocity**.

Problem Statement

In the digital age, the timeliness of accurate, accessible and actionable data provides a competitive advantage for all firms.

Strategically, most firms want to push master and transactional data assets into Data Lakes for analysis, insights generation and interrogation by AI and Machine Learning robots, to drive further process automation.

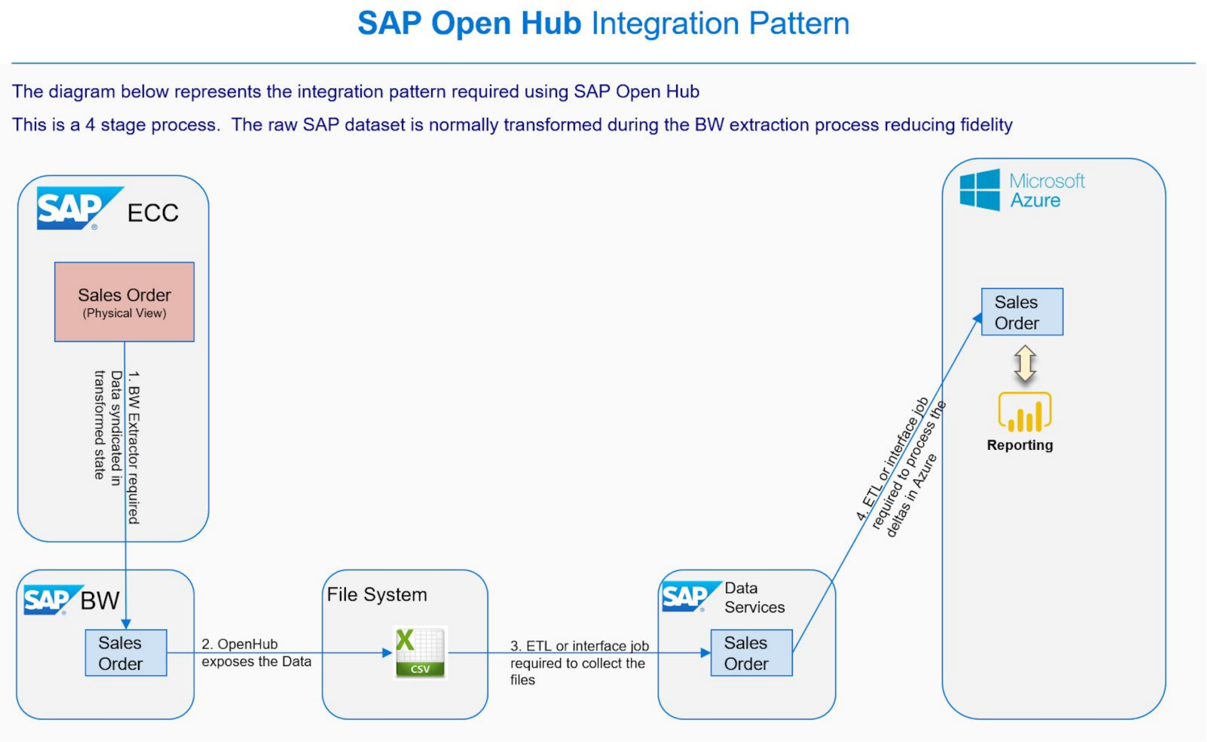
The typical integration pattern to ingest/syndicate data from SAP systems to Data Lakes relies on superfluous middleware application technologies to fulfil the data movement.

An example integration pattern is:

SAP-ECC -> SAP-BI/BW -> ETL (BODS/SDS) -> Data Lake (Azure)

The BW and ETL layers add little or no transformational value to the dataset and simply act as a forwarding/relay mechanism.

SAP also provide a solution called OpenHub which is designed to improve data integration opportunities with SAP via BW, however this further complicates the overall integration pattern and introduces additional delays in the data propagation process. See diagram below:



SAP Open Hub Integration Pattern

The diagram below represents the integration pattern required using SAP Open Hub

This is a 4 stage process. The raw SAP dataset is normally transformed during the BW extraction process reducing fidelity

SAP ECC
Sales Order (Physical View)
1. BW Extractor required
Data syndicated in transformed state
SAP BW
Sales Order
2. OpenHub exposes the Data
File System
CSV
3. ETL or interface job required to collect the files
SAP Data Services
Sales Order
4. ETL or interface job required to process the deltas in Azure
Microsoft Azure
Sales Order
Reporting

There are several problems with this approach some of which include:

- **Cost**

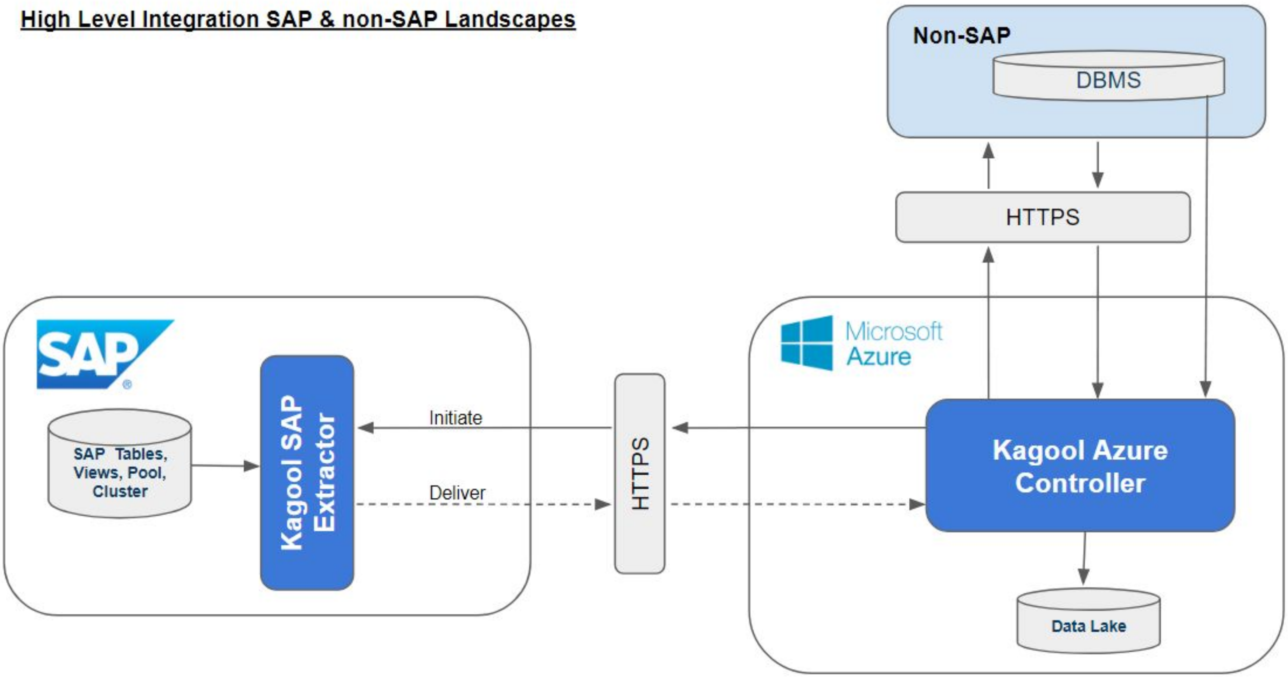
- Inflated administration/support costs
    - Unnecessary licencing costs for technologies not adding value in the process.
    - Multiple technical skill sets required to support and maintain
    - Excessive infrastructure required to support data movement
    - Significant development required just to add one additional table to the Lake
- **Complexity**
    - Too many different technologies involved with too many integration points
    - The incremental addition of just one table requires design, build and testing in many separate technologies
    - Excessive moving parts - more things to go wrong and impact overall data movement process. Multiple points for failure to occur.
    - Technical difficulties in managing large delta volumes
- **Timeliness**
    - Data will age by the time it reaches the lake and could be more than 24 hours old
    - Not normally possible to have near real-time Data Lake syndication in large enterprises
- **Security**
    - Many entry points for Man In The Middle attacks
    - Data often held at rest outside of the security of the ERP or Analytics platform
    - Data generally held/managed in an unencrypted state when outside the ERP or Analytics Platform
    - More technologies means many more users have access to the data in the support of this process, thus increasing risks

Velocity: Azure Data Lake Ingestion

The Kagool SAP -> Data Lake solution enables a direct connection between SAP System(s) and Data Lake technologies via Azure Cloud components, removing the need for any middleware layer(s) and optimising the data transfer delivery.

SAP-ECC ->~~-SAP-BI/BW -> ETL (BODS/SDS/Data Stage)~~ -> Data Lake

The Kagool solution is fast to deploy and involves installation on both ends of the connection (SAP Side and Data Lake side), once installed the movement of data is fully parameter controlled and managed by your administrators



New tables can be added to the Lake in literally **minutes with *no development required*.**

The management of data flows across the IT landscape into the Data Lake is managed by the Kagool Azure Controller which can be considered as a Data Orchestration solution. From here, administrators control the scope and delivery mechanism of how data flows into the Lake, which is performed through configuration.

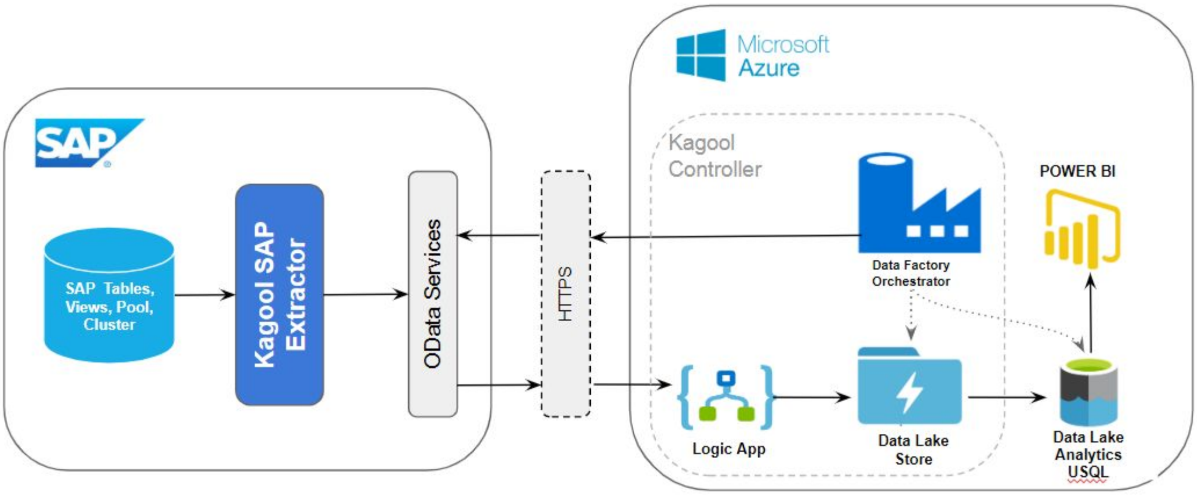
The Controller manages data syndication from any number of SAP systems (ECC, EWM, SRM, GTS etc) and any number of non-SAP systems whose connections can be via application or database layer at customer discretion.

The Controller holds information about which objects are in scope from which SAP systems and also includes ***how*** the data should be transmitted to the Lake eg:

- Recordset size
- # Parallel streams
- Compression technique

All parameters are fully configurable from within the Controller Interface by admins to fine tune interface at object level.

The diagram below provides further details on the SAP integration including the Azure components used within the current solution (Version 1.0). The next release will incorporate Data Bricks.



USPs

The Exec Summary of the Kagool solution USPs is summed up in the following words - *Faster, Cheaper, Simpler & More Predictable Outcomes*

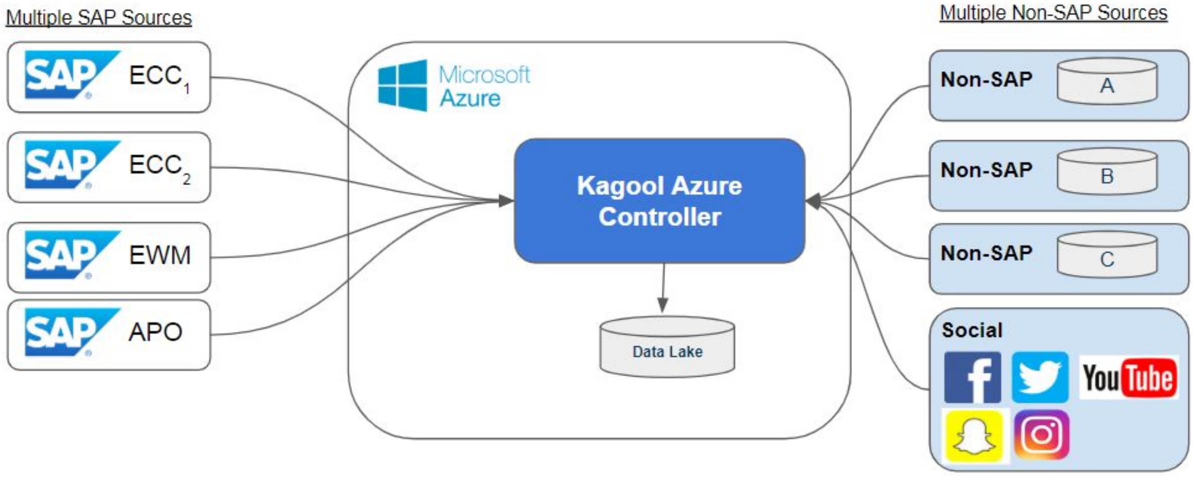
For a detailed description please see the bullets below:

- **No additional Infrastructure Required to Deploy**
  - the solution since it sits on client's existing SAP/Azure infrastructure
- **No New Underlying Technologies for Client's to Learn/Support**
  - it only uses native SAP and Azure technology
- **Makes full use of SAP and Azure Technical Resource Management Solutions**
  - does not introduce any further complexity
- **Fast to Install**
  - SAP Transports and Azure install packages provided
  - Deployable in days per environment
- **Faster to Roll out New SAP Data Object Integrations for Projects**
  - No build required, pure config
- **Projects Delivered Faster, Cheaper with more Predictability**
  - Fewer moving parts to change - smaller teams, less tasks to perform
  - Easy to commission new interfaces via config - no developments required
  - Performance testing can be quickly tuned/managed in production via config
- **Significant Ongoing Cost Savings**
  - No middleware technologies to maintain
  - No development required for new Interfaces nor interface changes
- **Simple Integration Pattern - Streamlined**
  - Fewer moving parts...Less things to go wrong
- **Source Schema Preservation**
  - Velocity preserves the source schema/metadata for downstream usage
- **Full CDC Delta Management**
  - Uses native SAP capabilities to identify delta record changes
  - Only the delta records are transferred to the lake
  - Delta records are fully merged into the target ADLS files without creating additional delta files
  - Does not follow Batch Management process unlike other solutions
- **Special SAP extractions**
  - SAP BW Extractors available as a source
  - SAP Reports available as a source
  - Executable t-Codes available as a source
  - SAP BW objects available as a source
- **Near Real-time Data Lake - Huge Opportunities**
  - Via constantly streamed data to the Lake
  - AI/ML processing using real-time data - instant reaction to events
  - Real-time analytics on global datasets available from the Lake

Solution Key Features

- The SAP Integration point is through the SAP application layer and not database layer. This ensures that SAP's automated internal controls for resource distribution (inc load balancing) & management are inherently part of this solution. SAP Environment and basis team(s) will also have full visibility and control of application resources
- Non-persisted SAP tables (view, pools, clusters) are fully resolved/translated via the Extractor
- **Control**
    - Admins can easily & instantly control the flow of data from SAP to Azure
        - Data Batch size
        - Parallelism
        - Compression
        - CDC type
    - Basis can control the system resources assigned to the Adapter within SAP
- **Configuration**
    - Object scope is managed via configuration in the Azure Controller, no development is required to add further tables to the Lake.
    - New SAP tables can be added to the scope for consumption into the Lake in minutes (via a new configuration entry), this could in theory be syndicated to the Lake in minutes.
        - Additional SAP Authorisations can be included to further control access.
- **Scalability**
    - The Kagool Data Lake integrated solution is fully scalable for any number of concurrent SAP systems and any number of Non-SAP systems within a Landscape including Social sources.

High Level Integration SAP & non-SAP Landscapes



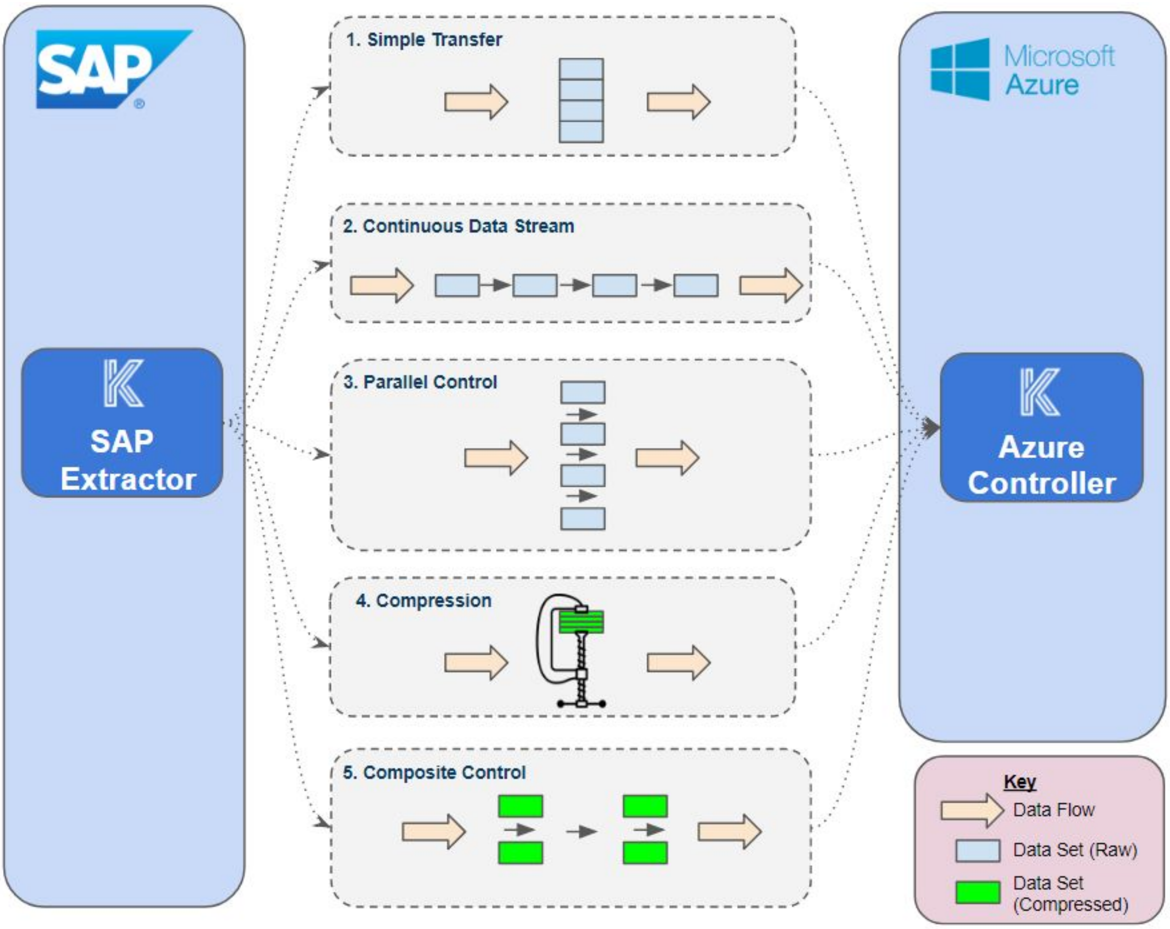
- **Timing**
    - By optimising the data integration patterns for your landscape objects into the Lake, time critical objects can be syndicated upon creation on a real-time basis enabling ML/AI processes in the Lake to immediately start processing the data.
    - The Lake can in theory be tuned to be as up to date as required based on object criticality

- **Cost Savings**
    - The incremental effort of adding a new SAP table to the Lake is the time it takes to add a new configuration entry to the Azure Controller (minutes).
    - The table could be syndicated into the lake minutes after saving the entry in the Controller - no development is required

Data Flow Controls

The data transfer is a highly sophisticated process, enabled for optimising data movement of very high volumes, extremely quickly and securely, meaning your Data Lake can hold a near real-time dataset - this can be critical for management decision making and also automated AI and Machine Learning (ML).

The way in which changed datasets (deltas) flow from SAP to the Data Lake can be finely tuned at individual data object level by user Admins from within the Azure Controller module. There are Pros and Cons for each method since there will be a processing overhead when applying intelligent syndication methods, therefore these may only be relevant for larger datasets.



Key methods include:

1. **Simple Transfer**
   - The full delta dataset is syndicated without any changes. This is useful for low volume delta processing
2. **Continuous Data Stream**
   - The delta record set is auto split into batches and streamed to Azure
   - This can be a continuous process running 24x7
3. **Parallel Control**
   - The delta dataset is auto split into a configurable number of parallel processes
   - This is useful for large datasets for faster transfer rates

4. **Compression**
    - Compression techniques can be applied to further improve data transfer rates
5. **Composite Control**
    - Azure Admins can use any combination of the Integration Control methods to finely tune the data flow for each object

Glossary

The table below outlines the key terms referenced in this document

Item	Description
POC	Proof Of Concept
AI	Artificial Intelligence
ML	Machine Learning
ETL	Extract Transform Load - Data Migration process
BODS/SDS/Data Stage	Examples of Data Migration Tools (Business Objects Data Services, SAP Data Services)
SAP BI/BW	Business Intelligence / Data Warehousing solution provided by SAP
SAP ECC	SAP Core resource Planning system - Finance, HR, Purchasing etc etc
SAP SRM	SAP System for Supplier Relationship Management
SAP EWM	SAP System for Enterprise Warehouse Management
SAP GTS	SAP System for Global Customs and Tax
SAP APO	SAP System for Supply Chain Planning
Syndication	The process of moving data across an IT landscape into different systems
Ingestion	The process of consuming large data sets into a Data Lake
Data Object	A set of data that can be grouped into a common noun for example Customers, Vendors, Purchase Orders, Invoices etc etc. Data Objects can be Master or Transactional Data
Data Compression	The process of reducing the file size of data without any loss of the data integrity. Data is transmitted once compressed.. Once the data is received a reverse process of decompression is applied to recover the data to the original state without any loss of integrity throughout the process
CDC/Delta	Change Data Capture - The process for identifying changes in data on a system between two different points in time
MS	Microsoft
USP	Unique Selling Proposition