RDX

BI ENVIRONMENT PLANNING GUIDE

Business Intelligence can involve a number of technologies and foster many opportunities for improving your business. This document serves as a guideline for planning strategies to begin successfully building an on-premises SQL Server or cloud-based Azure Business Intelligence environment.

Written By:
JEREMY FRYE

For more information:
rdx.com // info@rdx.com

SQL Server On Premise BI General Planning

Extract, Transform and Load (SSIS)

Extract, Transform and Load (ETL) describes the process of moving data between databases, tables, files, etc. with the ability to change and conform it to meet business requirements. Within the SQL Server platform, Integration Services (SSIS) is the design tool used to implement such a solution.

Here are some things to consider when planning for an on-premise SSIS environment:

- Memory
  - SSIS takes advantage of memory more than any other hardware component from a performance standpoint
  - Make sure you allocate enough memory on the server outside of what is allocated for the SQL Server database engine to allow as much caching of pipeline data as necessary
- Processors
  - SSIS is configured by default to execute the number of tasks in parallel equal to the number of logical processors plus 2; this setting can also be adjusted
- Framework
  - It's recommended to set up a framework for SSIS to use for audit and logging purposes as well as for scalability and migrations
  - Use configurations for 2008 R2 and Prior
  - 2012 and newer is built in using the SSIS catalog
- Design
  - Use variables when possible
  - Avoid too many parent/child package calls *(this can be a troubleshooting nightmare)*
  - Consider reusability
  - Based on the task(s), determine when to let SSIS do the processing and when the SQL database engine will be better
- Storage
  - SSIS 2008 R2 and older, use the package store for database and file system storage
  - SSIS 2012, use the catalog

Data Warehouse and Analysis (SSAS Tabular)

Analysis Services within the SQL Server tool stack is used to compile current and historical information usually in a denormalized schema, that allows for faster aggregations, analysis and data mining to help forecast business trends and help make better business decisions.

Here are some things to consider when planning for an on-premise SSAS environment:

- Memory
  - Unless you are using direct query, tabular SSAS models utilize an in-memory columnar storage mechanism called the Vertipaq engine. Data in the Vertipaq engine is also compressed
  - By default, SSAS total memory limit is set to an 80% threshold of physical memory or virtual address space of the memory on the server. Once this threshold is hit, processes will begin shutting down in order to deallocate memory aggressively
  - It's typically a best practice to separate the SSAS service and the database engine so they do not compete for resources
- Processors
  - Parallel processing is one of the keys in maintaining good performance in SSAS; because large amounts of data are being aggregated and rolled up, splitting this across as many processors as possible will help improve performance
- Disks (Multi-dimensional)
  - When talking about warehouses, you are typically talking about years of historical data
  - This data is stored in cubes, aggregations and object processing methods cached and stored on disk
  - Adequate disk space is essential for ever-growing warehouses
  - RAID 1/0 or RAID 5 configurations are typically ideal from performance standpoint when it comes to traditional hard disks
  - Solid state drives are much more preferable over traditional hard disks as they do not have mechanical parts but use NAND-based flash memory
- Data Model Design
  - Determine KPI metrics and goals
  - Establish a grain of fact *(Should only consist of measurable data)*
  - Realize what kind of schema will work best for your needs *(star, snowflake, etc.)*
  - Use surrogate keys when possible
  - Define aggregations
  - Logically partition semantic measures
  - Physically partition fact tables to align with logical semantic partitions
  - Use dimension hierarchies for better usability and drill-down functionality

Reporting (SSRS)

Reporting Services in the SQL Server tool stack is simply a graphical design tool to display your data in whatever output format desired. Analytical data in a data warehouse or transactional data sitting in an applications database system can be viewed in a number of ways with compelling visualizations.

Here are some things to consider when planning for an on-premise SSRS environment:

- Processor and Memory
  - The performance of reports is mainly driven by the SQL queries used in the reports; although visual basic expressions can be written and charts and graphs rendered, the processing effort is small
  - The report server databases can be hosted on any SQL instance but depending on the activity of the system, you may want to consider installing the report server instance on its own server
- Design
  - The main thing to consider here is query design; use efficient query writing techniques to avoid performance bottlenecks
  - Determine when to build data sets with logic in SQL code and when it may be easier or more efficient to write an expression
  - Depending on the complexity and frequency of the reports being run, as well as the activity on the source data server, it's sometimes best to replicate the source data to a read only reporting database or instance to alleviate I/O

Azure Cloud Architecture BI General Planning

Azure Data Factory

Azure Data Factory is a cloud-based software as a service that allows you to perform automated data consolidation, transformation and movement through pipelines and work flows. Azure Data Factory has two different versions. Version 1 is all code based using JSON programming. Version 2 utilizes a GUI for development. SSIS can also be integrated within a Data Factory v2 Pipeline by utilizing an Integration Services runtime component. This will allow SSIS developers to continue to leverage the integration services platform within a hosted environment.

Here are some things to consider when planning for an Azure Data Factory Implementation:

- Service Type
  - For a data factory instance, understand if you need a Data Pipeline or a SQL Server Integration service. There are price differences between the two
  - For SQL Server Integration service, if you already have a SQL License, you can save with an Azure Hybrid Benefit
- Data Factory Operations
  - In a data factory pipeline, you are charged by read/write and monitoring operations measured in entity units
  - Each entity unit consists of 50,000 entities
  - Entities are datasets, linked services, pipelines, integration runtime, and triggers
  - When building data factories, consider design strategies that can accomplish the data integration in as few entities as possible to cut down on cost
- Azure Integration Runtime
  - There are two types of pipeline activities. External activities are components that utilize architecture outside of Azure Data Factory. The execution costs of each type of pipeline activity is different
  - Understand what kind of activities you need to utilize in your implementation
  - The orchestration of the activity runs are measured in thousands
- Orchestration
  - Pipeline activities are run in orchestrations that can be automated
  - Orchestration costs are measured in thousands. Plan for the costs of running your pipeline activities by determining the schedules of your ETL/ELT solutions

Azure Databricks

Azure Databricks is a cloud-based software as a service used for streamlining workflows in a collaborative workspace. Databricks can be used in a myriad of ways in the modern data warehousing architecture. It is built on the Apache Spark analytics engine. It has the ability to do ETL and analytics using a number of popular developer languages. It can be integrated with other cloud service platforms seamlessly or as its own all in one data analytics and artificial intelligence solution.

Here are some things to consider when planning for an Azure Databricks Implementation:

- Instances/Billing Options
  - Determine if your solutions need to run on a steady basis and you need reserved capacity. You can save money over time by going with a yearly reserved option for virtual machines
  - There are two tier options *(Standard and Premium)*. Based on how much data and analytics you are performing, choosing the tier structure wisely can have a big impact on your monthly charges
  - The standard and premium tier instances give you a variety of options to choose from as it relates to cpu and memory.
- Workloads
  - Databrick units (DBU) are the units of measurement for processing capability per hour
  - Determine your level of data manipulation and analytics vs the need for controlling workflow by building and automating jobs to understand which workload to leverage
  - The DBU cost per hour is different based on the workload
- Integration/Collaboration
  - Utilize notebooks based on the language skill set of preference within your development teams
  - Understand your change control process

Azure Data Lake

Azure Data Lake is another option for storage and analytics of big data. Like Databricks, it is also a cloud-based software as a service. Azure Data Lake is built on the HDInsight platform which consists of Apache Spark and Hadoop clusters for dynamic scale and performance. Data Lake Storage and Data Lake Analytics make up the components of Azure Data Lake. Data Lake Storage is Azure Blob storage optimized for analytics workloads. Data Lake Analytics allows data scientists, developers and business professionals to easily develop, run and scale massively parallel data transformation and processing programs in U-SQL, R, Python, and .NET over petabytes of data. With no infrastructure to manage, you can process data on demand, scale instantly, and only pay per job.

Here are some things to consider when planning for an Azure Data Lake Implementation:

- Storage
  - The amount of storage used or committed to determines your charges in Azure Blob Storage. Even though you can dynamically scale, have an idea of the size of your data to help project costs
  - You are also charged by read and write transactions called Transaction Units. 1 Transaction Unit includes 10,000 transactions
- Analytics
  - Analytic Units are the unit of measurement for Data Lake Analytics. These are units of computation for jobs. Analytic Units gives the jobs access to a set of resources for CPU and memory
  - 1 Analytic Unit is equal to 2 CPU cores and 6 GB RAM. This could be subject to change as Microsoft sees fit in the future
- Usage/Use Case
  - Since there is some overlap in functionality around storage and analytics between Azure Data Lake and Azure Databricks, understand your budget, business needs and BI staff expertise.
  - Azure Databricks charges you based on the workload, tier and instance you select. Azure Data Lake components are based on storage used, read/write transactions and data analytic units
  - Language skill set, real-time interaction, notebook collaboration and scalable runtimes are a few things to consider when determining to use Data Lake Analytics or Databricks.

Azure SQL Data Warehouse

Azure SQL Data Warehouse is a Massively Parallel Processing (MPP) cloud-based enterprise data warehouse. It used for big data applications and analytics over larger sets of data. Data is typically fed into Azure SQL Data Warehouse with PolyBase T-SQL. The SQL Data Warehouse stores data in columnar storage and can be scaled in minutes. Data is then fed to visualization tiers such as Power BI or even used with Azure Analysis Services.

Here are some things to consider when planning for an Azure SQL Data Warehouse Implementation:

- Performance Tier
  - Pricing is based on compute and storage. Compute is measured in Data Warehouse Unit (DWU) Blocks. A DWU is CPU, memory and IO bundled together
  - Increasing DWUs can give better performance of scans, aggregations, CTAS statements, PolyBase load operations and concurrent queries
- Storage
  - Pricing is per Terabyte (TB) and not transactions
  - Storage size includes the size of the data warehouse and 7 days of incremental snapshot storage
- Architecture Planning
  - Azure SQL Data Warehouse can be used with a variety of other SaaS tools and architectures. Understand the business needs, size/growth of data and scalability options to know how to involve Azure SQL Data Warehouse appropriately
  - Use Azure SQL Data Warehouse with Azure Analysis Services for a modern data warehouse to integrate scale, performance and isolation of resource and security
  - Utilize Azure Databricks for analysis and transformation of structured and unstructured data to feed Azure SQL Data Warehouse
  - Migrate your existing on-premises data warehouse to Azure SQL Data Warehouse

Azure Analysis Services

Azure Analysis Services is a cloud-based Platform-as-a-Service (PaaS) solution for enterprise-level data modeling. Because it is a full managed service, pricing and performance is based on tier and instance choice, but the functionality of managing a semantic data model is equivalent to the way SQL Server Analysis Services works.

Power BI

Power BI is an interactive business and analytics visualization tool. It allows you to develop insights and answer questions about your business by connecting and integrating a variety of data sources to present in dashboards and reports or even embed in applications and websites.

Here are some things to consider when planning for a Power BI environment:

- User Capacity
  - There are different licensing structures for Power BI. Power BI Desktop can be used for free by anyone at no cost; however, sharing, security and deployment options are limited and not ideal for an enterprise solution
  - Power BI Pro gives more flexibility with the PowerBI.com. It enables you to develop and author, collaborate and share, and audit and govern data, reports and dashboards
  - Power BI Premium is for larger capacity planning. You are able to share with users inside and outside of the organization, manage and control dedicated server resources, have access to more refresh and data sizing options, and provide an on-premises implementation with Power BI Report Server
- Data Management and Governance
  - Decide what kind of deployment works best for your organization. IT-Managed or Corporate BI approaches usually work best
  - Create workspaces among the appropriate development teams to maintain a level of security and authorization and control of reporting content
  - Build apps that define the source of truth among various lines of business for governed report development and for easier maintenance of cleaned data
  - Document your data models and report logic by creating data dictionaries for your data sources and utilizing DMVs and tabular model metadata
- Data Modeling and Visualization
  - Integrate an enterprise data warehouse for streamlining data sets that can facilitate many avenues of reporting throughout the organization
  - Leverage the power of M code to transform and model data sets within the PBIX file when a traditional data warehouse schema when ETL is not available
  - Utilize DAX for complex calculations that can't be satisfied by general aggregations based on the model design
  - Use visualizations that can answer questions or tell a story about the data appropriately
  - Find custom visualization from the Market Place when standard visualizations don't seem to apply for the use case
  - Allow dashboards to show summarization of data at the highest level and utilize drill through or drill-down capabilities for more line item data exploration

About the Author: Jeremy Frye

Jeremy Frye is RDX's Business Intelligence and Data Warehousing Team Manager. During his nine-year tenure at RDX, Jeremy has held both technical and management roles and has an extensive background in business intelligence architecture, development and SQL Server. As an active member of the Professional Association for SQL Server (PASS) community, Jeremy frequently participates in local Microsoft BI user group meetings and speaks at SQL Saturdays throughout the United States. He also hosted a speaking session at the 2018 PASS Summit, the largest annual conference for SQL Server and Microsoft BI professionals. In his free time, Jeremy likes to work out, listen to music and spend time with his family.

About Remote DBA Experts, LLC

Founded in 1994, RDX is one of the largest independent providers of database infrastructure and cloud management services. RDX provides managed services to hundreds of clients, both on premise and in the cloud. RDX supports a wide range of cloud environments including Microsoft Azure, Amazon AWS, and Oracle DB Cloud; database environments including Oracle, SQL Server, MYSQL, PostgreSQL, DB2 and MongoDB; and operating systems including Windows and all major UNIX/Linux offerings. For over two decades, RDX has helped hundreds of organizations lower support costs while increasing data infrastructure performance and availability. RDX's expert staff of highly-trained professionals is backed by a monitoring and support infrastructure that has been continuously improved and enhanced throughout its history. More information is available at https://www.rdx.com.