BI Cloud Architecture in Azure

Are you stuck working with ETL instead of actually improving your business with data analytics?

Do you want to get access to great scalable computing power for less money than ever?

Random Forest have been building PaaS (Platform as a Service) solutions in Azure for a couple of years and due to the increased number of tools available in Azure we have established a best practice for building an Azure data solution.

Why PaaS?

Going from a traditional server structure to a PaaS based solution is a game changer for BI. This dramatically reduces the need for setting up and maintaining the platform and outsources the infrastructure. Usage of PaaS is paid by the hour and enables on demand scaling and disaster recovery through geo-redundancy. The service ensures you are always on the latest version and the licensing is included which minimizes the total cost.

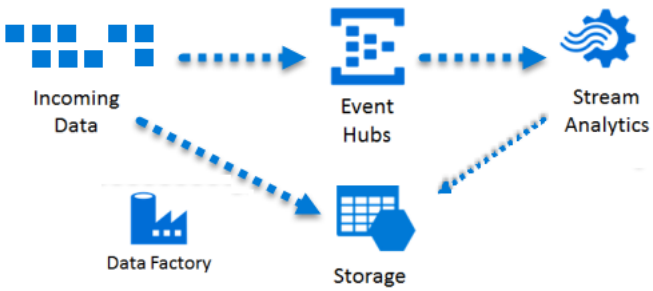
Moving to Azure opens up new possibilities which raises the need for a different design pattern. The cloud challenges old architectures and requires a new way of thinking. Storing data is no longer a cost driver, data streaming is now preferred over batch loading and the toolkit is vastly expanded.

Sourcing

Pushing data to the Event Hub creates a standardized way of receiving different kind of data with minimum time delay. Stream Analytics is used as one of the Event Hub subscribers. This enables you to show real time data in Power BI and save the data in Azure Storage for data warehouse staging or a data lake.

If you need to fetch data on a regular basis we recommend using the Data Factory to load data from any source, internal or external, into Azure Storage for data warehouse staging or a data lake. In case you can't allow the usage of a data factory gateway and have a legacy SSIS server you can also use that to push data to Azure Storage.

The blob storage in Azure Storage is preferred since it is easier to use and is more flexible with data types. Having the data in Azure Storage removes the need for a staging environment, since it can store endless amount of data and is accessible through PolyBase. The cost for storage is minimal.



The Data Warehouse

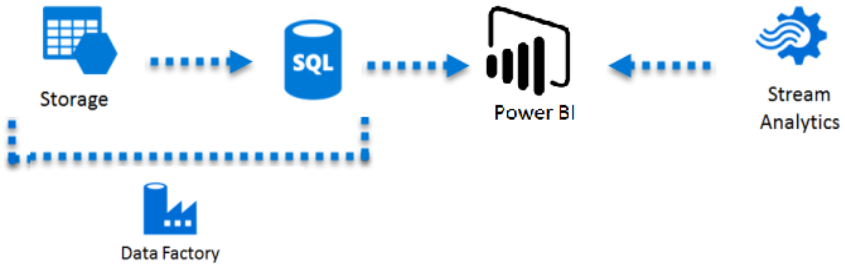
We recommend setting up an architecture that supports using both Azure SQL Database and Azure SQL Data Warehouse. The two are similar but have different strengths and capabilities. Having a design pattern that works for both enables us to move the solution more easily when needed.

Random Forest AB
Kornhamnstorg 49
Stockholm
Sweden

Version: 1.0
Created: 2016-10-19SE-111 27
Contact: Richard Lautmann
Phone: +46 (0) 70 281 01 68

As a general principle, you should consider using Azure SQL Data Warehouse when the data volume is greater than 500 GB or there is a need for heavy data analytics. If you have many concurrent request on smaller volumes with a need for quick response times, we recommend using the Azure SQL Database.

Getting the data from the Azure Storage into the database can be achieved by using PolyBase in stored procedures, orchestrated by Data Factory. A key concept in Data Factory is slicing the data by time, enabling parallel processing and rerunning individual slices. Real time data should be handled in Stream Analytics and the output can also be pushed to Azure Storage for archiving and usage in the data warehouse. This also applies for Machine Learning outputs.



We recommend having a metadata driven approach to the loading process with auto-generating stored procedures, separating business logic from loading. Removing lookups and enabling parallel loading by using hashed business keys instead of integer sequence column reduces complexity and increases performance. Hashed keys remain the same in all environments and helps when moving data between production and test environment.

Choosing how to model your data warehouse depends on your preference. Data Vault and Anchor Modeling gives a highly flexible model but increases complexity while a more traditional normalized data model is easier to get started with but less flexible.

Columnstore Indexes should be used, especially for big tables, removing the need for creating aggregate tables.

Visualizations and analytics

Power BI is a great tool for visualizing and analyzing data that fits very well in the Azure stack and can even be embedded in custom web applications. For more advanced and operational analytics you can use Azure Machine Learning and R.

Summary

By following this design pattern focus is moved to business logic instead of handling infrastructure, load logic, performance enhancements, licensing etc. Our experience is that you also save considerable time on ETL development with simpler and more specialized Azure services. You can be cost effective and enhance performance by scaling your solution based on the current need.

Consult us for your next step into the cloud!

Authors: Richard Lautmann & Gösta Boström