Building a small data warehouse in a big data world

Helen Gore, Iridium Insights

SQL Relay
October 2017
FEEDBACK

SQLRelay.co.uk/feedback

Live Now
Win a Lego BB8!
Helen Gore:
• 17 years of experience building data warehouses with SQL Server and SSAS
• CTO and BI Consultant at Iridium Insights – a data analytics consultancy in Bath.

Why am I doing this talk?!
• Fast pace of change
• Big data tools
• Cloud options
• Help to navigate options

“Small data warehouse”
Small production environment
Proof of Concepts
Demos
Data mart
Main objective: To help identify options to evaluate for your next project

<table>
<thead>
<tr>
<th>Agenda</th>
</tr>
</thead>
</table>
| • Review different components in turn  
  ➢ On-prem and Cloud  
  ➢ Pricing  
  ➢ Pros and Cons  
| • What components work well together?  
| • Criteria that may affect your choices  
| • Interactive tool  |
Assumptions

Some experience with:
- Fact and dimensional modelling (star schemas)
- SSAS

An awareness of:
- Power BI
- Azure

- Nightly/scheduled refresh – haven’t considered real
time/live streaming of data

- New projects – not migration of established DWs
Options to cover

On Premises

- Power BI
- SSAS Multidimensional
- SSAS Tabular

Azure

- Azure SQL Database
- Azure SQL Data Warehouse
- Azure Analysis Services
- Azure Data Lake Store
- Azure Data Lake Analytics
- Azure Virtual Machine

Big Data tools
• Steep learning curve to get started
• Overkill for simple models
• MDX is difficult to learn
• Some issues if importing data to Power BI
• Good if very large complex systems

Microsoft say:

and deployed. The bulk of this topic compares these two types so that you can identify the right approach for you.

For new projects, we generally recommend tabular models. Tabular models are faster to design, test, and deploy; and will work better with the latest self-service BI applications and cloud services from Microsoft.

Overview of modeling types
• Max model size is 1Gb
• Scalable?
• Hard to track ETL code - no version control for M
• Some ETL tasks are easier to achieve in a SQL environment
• Difficult to collaborate

✓ Tie in with data viz platform makes for simple dev experience

✓ Free
Power BI - used to replace Tabular

- Max model size is 1Gb
- Scalable - convert PowerBI model to SSAS Tabular
- Can track ETL in source control
- New fields from SQL brought in automatically on refresh
- One less step to maintain
SSAS Tabular (on-prem)

- Only Tabular model – no multidimensional
- Scale up and down as needed
- Have to stop server to scale
- Can pause if not in use to reduce costs
- Online web designer for small changes – lose source control!

Both:
- Same features in both environments
- Quick to get started
- Great governance and security
- Developer interface is clunky (Visual Studio)
- Lots of investment from MS

- Can migrate model from on-prem to Azure when needed

Azure Analysis Services

- Only Tabular model – no multidimensional
- Scale up and down as needed
- Have to stop server to scale
- Can pause if not in use to reduce costs
- Online web designer for small changes – lose source control!

£73 a month for Dev tier (can be used in production)
£238+ a month Basic tier
£450+ a month Standard tier

If more than 1Gb data...
If need front end that is not Power BI...
Azure SQL Database

- DBaaS – no SQL license required
- Can connect via SSIS and Management Studio
- Data Factory can schedule stored procs
- Azure SQL Analytics - monitoring
- No SQL Server Agent, Cross-DB queries, CLR

From £3.72 a month (for 5 DTUs)

- May need more DTUs to get sensible performance
- Can scale the DTUs/pricing up and down whilst the database stays live
- Can’t pause – always paying at least base price
- Up to 4TB (if on premium pricing)

Coming soon...

Managed Instance

- In preview
- Closer to on-prem SQL experience
- Will include Instances, SQL Server Agent, Cross-DB queries, CLR, R Services
Azure SQL Data Warehouse

- DWaaS – no SQL license required
- Significant investment in development time to configure it correctly for your data
- Min of 1TB before should consider using it

Storage costs from £100 per month (1Tb minimum Premium storage)
Compute costs range from £770 per month to £140,000 per month

- Although can pause compute, it must be on for further tools to connect to the data
Azure Virtual Machine

- Same options as an on-prem solution including SSIS, SSRS, SSAS MD, etc
- Can back up the SQL Database directly from Azure without being in the VM
- Can stop server when not in use at zero cost
- Can schedule stopping server every night

- VM with Standard SQL License (8Gb RAM) ~ £400 a month
- Windows only VM, install SQL with own license - from £36 a month
Azure Data Lake

Data Lake Store

- Storage 3p per GB per month.
- Compute 2.5p a minute per AU
- Transaction charges (minimal)

E.g. 10GB + 0.03p X 12 months = £3.60 a year for storage
~£3.60 a year for transaction charges
10 min nightly job = £91 a year for compute

- Compute is pay-as-you-go (pay for time job is running)
- Can access storage when compute not running (unlike Azure SQL DW)
- Storage scales instantly – no limits on data volumes!
- Now has connector to Azure Analysis Services
- Can import data to Power BI (blog)
- U-SQL fairly straightforward for T-SQL users
- Scheduling in Data Factory / Azure PowerShell quite techie

Data Lake Analytics

- Examples are not focussed on DW models as Microsoft designed the Data Lake to act alongside the DW, not to be the DW

Visual Studio Emulator – free local version
Data Lake possible architecture

- SQL Data Source
- Multiple csv files
- PowerShell script to export data to one csv file per table
- Upload csv files to Azure Data Lake Store
- Use U-SQL to extract data from csv files into an ADLA database
- Use U-SQL to transform data into fact and dimension tables
- Output fact and dimension tables to csv files
- Power acts as Tabular model and front end Data Viz tool
- Data Lake Analytics

Data Lake possible architecture

SQL Data Source

Multiple csv files

PowerShell script to export data to one csv file per table

Upload csv files to Azure Data Lake Store

Use U-SQL to extract data from csv files into an ADLA database

Use U-SQL to transform data into fact and dimension tables

Output fact and dimension tables to csv files

Input to Azure Analysis Services

Import data into Azure Analysis Services

OR...

Power BI only acting as front end with live connection to AAS
Component combinations

Data Source → ETL method → Star Schema → SSAS → Front end
## Options for Star Schema Platform

<table>
<thead>
<tr>
<th>Platform</th>
<th>Costs</th>
<th>Data Size</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power BI</td>
<td>Free</td>
<td>Up to 1Gb</td>
<td>No</td>
</tr>
<tr>
<td>SQL Server</td>
<td>SQL License and hardware</td>
<td>Depends on spec</td>
<td>No</td>
</tr>
<tr>
<td>Azure SQL Database</td>
<td>From £3.72 per month</td>
<td>Up to 4Tb</td>
<td>Yes</td>
</tr>
<tr>
<td>Azure SQL Data Warehouse</td>
<td>From £770 per month</td>
<td>From 1Tb</td>
<td>Yes</td>
</tr>
<tr>
<td>Azure Data Lake</td>
<td>Storage 3p per Gb per month, Compute 2.5p per minute</td>
<td>Unlimited</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Options for SSAS platform

<table>
<thead>
<tr>
<th>Platform</th>
<th>Costs</th>
<th>Data Size</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power BI</td>
<td>Free</td>
<td>Up to 1Gb</td>
<td>No</td>
</tr>
<tr>
<td>SSAS Multidimensional</td>
<td>SQL License and hardware</td>
<td>Over 1Tb recommended</td>
<td>No</td>
</tr>
<tr>
<td>SSAS Tabular (on-prem)</td>
<td>SQL License and hardware</td>
<td>Up to 1Tb recommended</td>
<td>No</td>
</tr>
<tr>
<td>Azure Analysis Services</td>
<td>£73 a month for Dev tier</td>
<td>Any</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>£238+ a month Basic tier</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>£450+ a month Standard tier</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Source:** Star Schema

**ETL method:**

- **SSAS Multidimensional:** SQL License and hardware
- **SSAS Tabular (on-prem):** SQL License and hardware
- **Azure Analysis Services:**
  - £73 a month for Dev tier
  - £238+ a month Basic tier
  - £450+ a month Standard tier

**Cloud:**

- Power BI, SSAS Multidimensional, SSAS Tabular (on-prem): No
- Azure Analysis Services: Yes
How do you start to decide which combinations may be best for your project?

Which criteria are important in making the decision?

How do you relate those criteria to the options available?
Project considerations

- Data Size
- Cost
- Front End
  - Skill set and development effort
  - Source control
  - Scalability and future proofing
  - Governance vs Self Service
  - Cloud strategy
Considerations – Data Size

Data Size

Cost

Front End

< 1Gb

1Gb +

1Tb +

Power BI “import data” models have 1Gb limit
Considerations – Costs

Data Size
- FREE
- LOW: < £500
- MED: £500-£5,000
- HIGH: £5,000 +

Cost

Front End

Ball park annual cost
Considerations – Front end

Power BI
- Import data
- Live Connection
- Direct Query

Range
- Any reporting tool that connects to SQL, SSAS and Azure Data Lake etc
What is your data size?

Do you want to use the cloud?

What is your budget?

What front end will be used?

Do you have a SQL License?

Possible Combinations
MS BI Selector (built in Power BI)
Available online at https://iridium-insights.com/insights/microsoft-bi-selector/

<table>
<thead>
<tr>
<th>Star Schema Element</th>
<th>SSAS Element</th>
<th>Front End</th>
<th>Cost</th>
<th>Cloud usage</th>
<th>SQL License Required</th>
<th>Scalability (to next data size tier)</th>
<th>Dev Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL Server</td>
<td>SSAS Tabular</td>
<td>Range</td>
<td>Medium</td>
<td>On-prem</td>
<td>Yes</td>
<td>Migration required</td>
<td>Medium</td>
</tr>
<tr>
<td>Azure Data Lake</td>
<td>Azure Analysis Services</td>
<td>Range</td>
<td>Medium</td>
<td>Cloud</td>
<td>No</td>
<td>Can scale up and down</td>
<td>High</td>
</tr>
<tr>
<td>VM</td>
<td>VM</td>
<td>Range</td>
<td>Medium</td>
<td>Cloud</td>
<td>Yes</td>
<td>Not possible</td>
<td>Low</td>
</tr>
<tr>
<td>Azure SQL Database</td>
<td>Azure Analysis Services</td>
<td>Power BI</td>
<td>Medium</td>
<td>Cloud</td>
<td>No</td>
<td>Not possible</td>
<td>Medium</td>
</tr>
<tr>
<td>SQL Server</td>
<td>Azure Analysis Services</td>
<td>Range</td>
<td>Medium</td>
<td>Cloud</td>
<td>Yes</td>
<td>Can scale up and down</td>
<td>Medium</td>
</tr>
</tbody>
</table>
• SSAS Multidimensional and Azure SQL Data Warehouse not for small data sets

• If no SQL license then Azure SQL Database may provide a suitable solution

• Whilst data lakes have been designed with big data in mind, the Azure Data Lake can be used to provide a low cost cloud solution for a simple DW or star schema.

• If have a SQL license would still recommend using SQL Server for the ETL and star schema in SQL, with either Power BI or SSAS Tabular as required
Resources

- Slides: [iridium-insights.com/insights/SQLRelay-slides/](iridium-insights.com/insights/SQLRelay-slides/)
- MS BI Selector [iridium-insights.com/insights/microsoft-bi-selector/](iridium-insights.com/insights/microsoft-bi-selector/)
- Big Data LDN – London Olympia 15/16 Nov – Free entry [bigdataldn.com/](bigdataldn.com/)
- Twitter: @helenfgore
- Email: Helen.gore@iridium-insights.com

SQL Bristol Meetup [https://www.meetup.com/SQLBristol/](https://www.meetup.com/SQLBristol/)

Next meeting: Thurs 2\textsuperscript{nd} Nov from 5.30pm