Swimming in the Data Lake

Presented by Warner Chaves

Moderated by Sander Stad

Thank You

microsoft.com

Empower users with new insights through familiar tools while balancing the need for IT to monitor and manage user created content. Deliver access to all data types across structured and unstructured sources.

hortonworks.com

Hortonworks develops, distributes and supports the only 100% open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise grade deployments. It is the only Hadoop-based platform available on both Linux and Windows.

aws.amazon.com

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale Microsoft SQL Server databases in the cloud.

red-gate.com

Redgate makes ingeniously simple tools for Microsoft technology professionals working with SQL Server, .NET, Visual Studio, Azure, TFS. Trusted by 91% of the Fortune 100.

JOIN PASS

PASS is a not-for-profit organization which offers year-round learning opportunities to data professionals

Membership is free, join today at **www.sqlpass.org**

Access to online training and content

Enjoy discounted event rates

Join Local Chapters and Virtual Chapters

Get advance notice of member exclusives

PASS

BIO

- DBA and consultant for 10 years.
- Previously an L3 DBA at HP in Costa Rica, now a Principal Consultant at Pythian in Ottawa Ontario.
- SQL Server MCM and MVP.
- Email: warner@sqlturbo.com
- Blog: SQLTurbo.com
- Company: Pythian.com

MVP
Microsoft®
Most Valuable Professional

@warchav

PASS

Agenda

Goal of today: provide an overview of the Azure Data Lake Service.

- Azure Data Lake Service
- Azure Data Lake Store
- Azure Data Lake Analytics
- U-SQL
- Demo
- Recap

5

What is it?

- It’s a cloud PaaS offering for Big Data.
- Management minimized.
- Can fit into an organizations larger data architecture.
- Can receive files with no specified schema.
- Can interface with the other analytics services like Machine Learning or Power Bi.

6

Data Lake not Data Landfill!

7

The Do’s and Don’ts

- Just because you can store anything, doesn’t mean you should.
- Remember there are still other pieces of the puzzle for handling data:
  - Azure SQL DB (OLTP or small DW)
  - Azure SQL Data Warehouse
  - DocumentDb or equivalents

- IT governance, control and procedures are still important, just how they are in your SQL Server.

8

Azure Data Lake Service

Data Store

- No limits on object size.
- No limits on lake size.
- Optimized for parallel read scans, low latency writes.

Data Analytics

- Aims to decrease the friction from a SQL developer to a big data developer.
- Provides U-SQL: SQL + C#
- Can run queries and jobs and scale up and down per job.

9

Data Lake Store

- An evolution of the regular Blob storage with no size limits.
- It’s just a file system in the cloud. The characteristics of the file system is what makes it special.
- It’s a scale out replicated file system like Hadoop with large files split over multiple machines.
- Compatible with Hadoop tools because it’s compatible with HDFS and exposes a WebHDFS REST API.

10

Data Lake Store Security

- Can integrate with Azure AD.
- Role based access can be done with groups.
- There are service level roles: Owner, Contributor, Reader, User Access Administrator.
- ACLs are allowed on the file system as well (only at the root during the preview).

11

Data Lake Store Ecosystem

Moving data

- Azure Data Factory
- Powershell
- Portal
- Apache Sqoop
- Apache Storm
- Azure Event Hub
- Azure Import/Export service

Processing

- Azure Data Lake Analytics
- HDInsight:
- MapReduce
- HBASE
- Storm

Visualizing

- Move analytics results to a relational store and display with any BI tool.
- Power Bi can connect and import data directly.

12

Data Lake Store Billing Model

USAGE	PRICE (PREVIEW)
Data stored	$0.04/GB per month
Data Lake transactions	$0.07/million transactions

- Each transaction is billable as a unit of 128KB or less. Larger transactions are billed in multiples of 128KB. For example, 1KB is billed as one transaction but 140KB is billed as two transactions.

13

Data Lake Analytics

- A PaaS service for processing Big Data.
- The sibling to Data Lake Store.
- Designed for scale out and parallelism.
- Designed to be easy to pick up for SQL professionals.

14

Running analytics

- Focus only on running jobs, not on infrastructure.
- The system can scale per job using Analytic Units for compute.
- Authoring is integrated with Visual Studio.
- The analytics can also integrated with other Azure services like SQL Db, SQL DW or regular blob storage.

15

U-SQL

- The query language of Data Lake Analytics.
- Mixes SQL semantics with the power of C#.
- Works with unstructured data by applying schema on read.
- Runs on a distributed parallel runtime.
- The data types are the C# types and are .NET objects.

16

U-SQL

- The query language of Data Lake Analytics.
- Mixes SQL semantics with the power of C#.
- Works with unstructured data by applying schema on read or structured data on tables.
- Runs on a distributed parallel runtime.
- The data types are the C# types and are .NET objects.

17

U-SQL Example

```
DECLARE @outpref string = "/output/Searchlog-aggregation";
DECLARE @out1 string = @outpref+"_agg.csv";

@searchlog = EXTRACT
UserId int, Start DateTime, Region string, Query string,
Duration int?, Urls string, ClickedUrls string
FROM "/Samples/Data/SearchLog.tsv"
USING Extractors.Tsv();

@rs1 = SELECT Region, SUM(Duration) AS TotalDuration
FROM @searchlog
GROUP BY Region;

OUTPUT @rs1 TO @out1 ORDER BY TotalDuration DESC USING Outputters.Csv();
```


18

U-SQL Catalog

- Relational storage and metadata inside the Data Lake Store.
- Databases – Has the familiar T-SQL-like Use command.
- Tables, indexes and stats – allows for partitioning and different distribution methods.
- Views
- Functions (in C# optionally).
- Stored Procedures.
- User-defined operators and aggregators (C#).

19

Service management components

- Users: who can work with the service.
- Data Sources: what data the jobs access.
- Jobs: what U-SQL to run and how.
- U-SQL Catalog: used to structure data and code so they can be shared by U-SQL scripts .

20

Data Lake Analytics Billing Model

USAGE	PRICE (PREVIEW)
Analytics Unit	$0.017 / Minute
Completed Job	$0.025 / Job

Each Azure Data Lake Analytics account has configurable quotas limiting the number of AUs that can be assigned to jobs and the number of concurrent jobs.

UNIT	LIMIT
# of AU per job	Up to 50 *
# of concurrent jobs per account	Up to 3 jobs *

21

DEMO

Recap

- Azure Data Lake is a big data PaaS offering.
- Just because you can store anything doesn't mean you should.
- The service is split into Data Lake Store and Data Lake Analytics.
- The U-SQL jobs make it easy for a DBA to work with it.
- Each one is billed separately and control one part of the lake equation.
- The Data Lake service can easily fit into a modern organization data architecture by filling the gap of a fully managed big data offering.