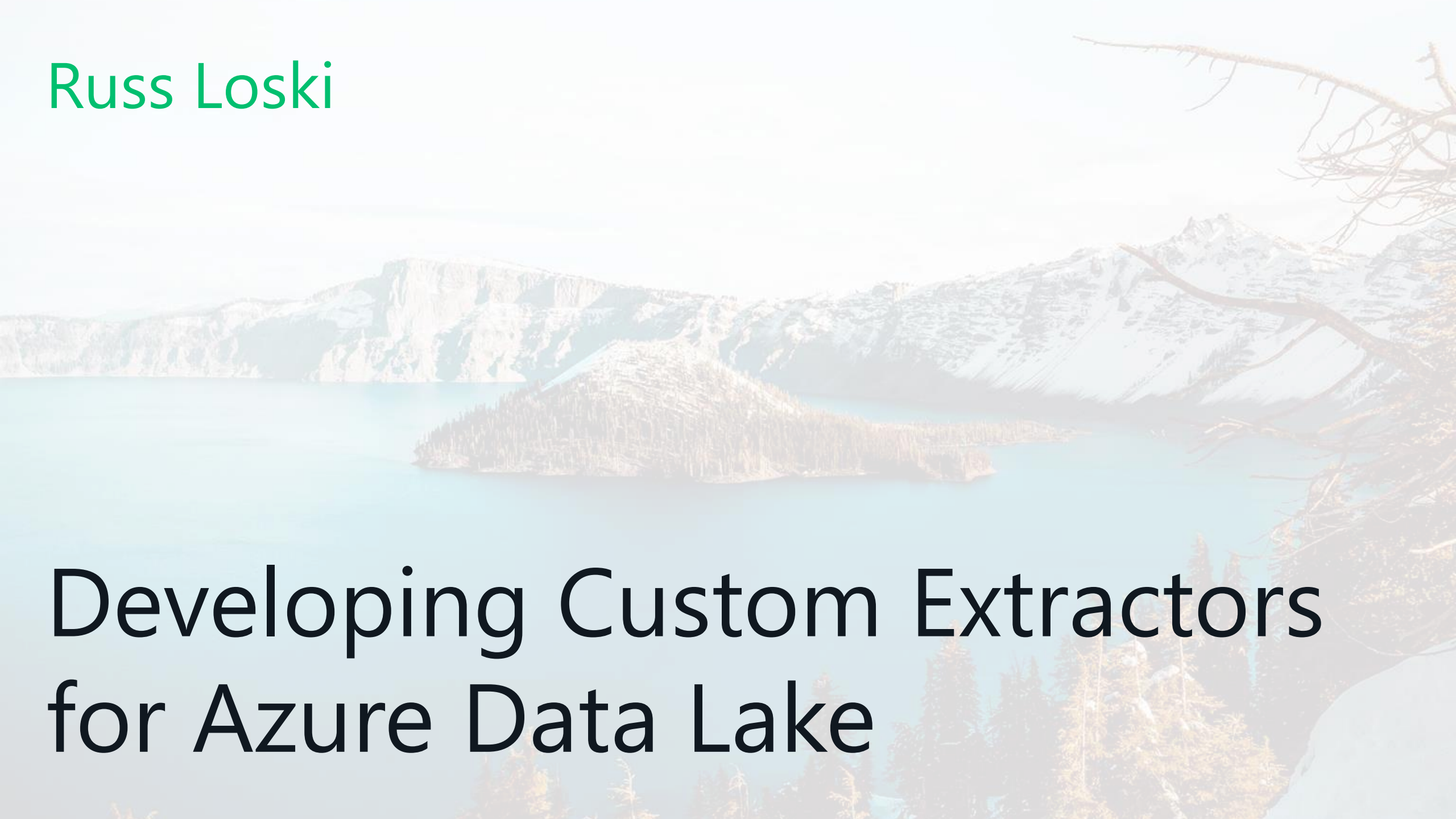
Russ Loski



Developing Custom Extractors for Azure Data Lake

Russ Loski

SQL Server ETL developer from Dallas Fort Worth

North Texas SQL Server Users Group

Grand dad for active 7 year old

Curious about data

Russ Loski

RussLoski@SQLMovers.com

www.SQLMovers.com

@sqlmovers

https://www.linkedin.com/in/russloski

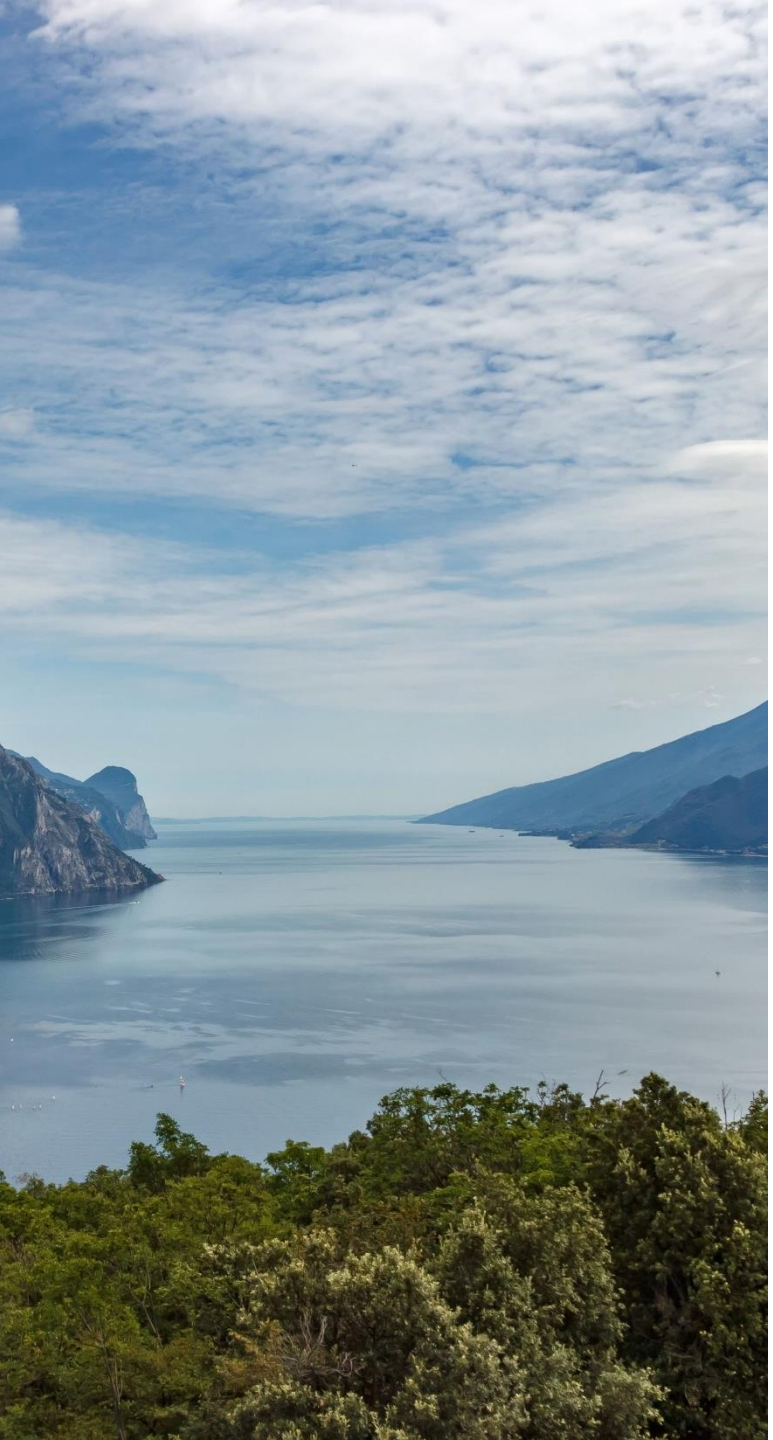
Agenda

What is a data lake?

What are U-SQL Extractors?

JSON Extractor

Flexible Extractor



Data Lake

What is it?

How do we use it?

Who is it for?

Data explosion

“1. The data volumes are exploding, more data has been created in the past two years than in the entire previous history of the human race.”

“20. And one of my favourite facts: At the moment less than 0.5% of all data is ever analysed and used, just imagine the potential here.”

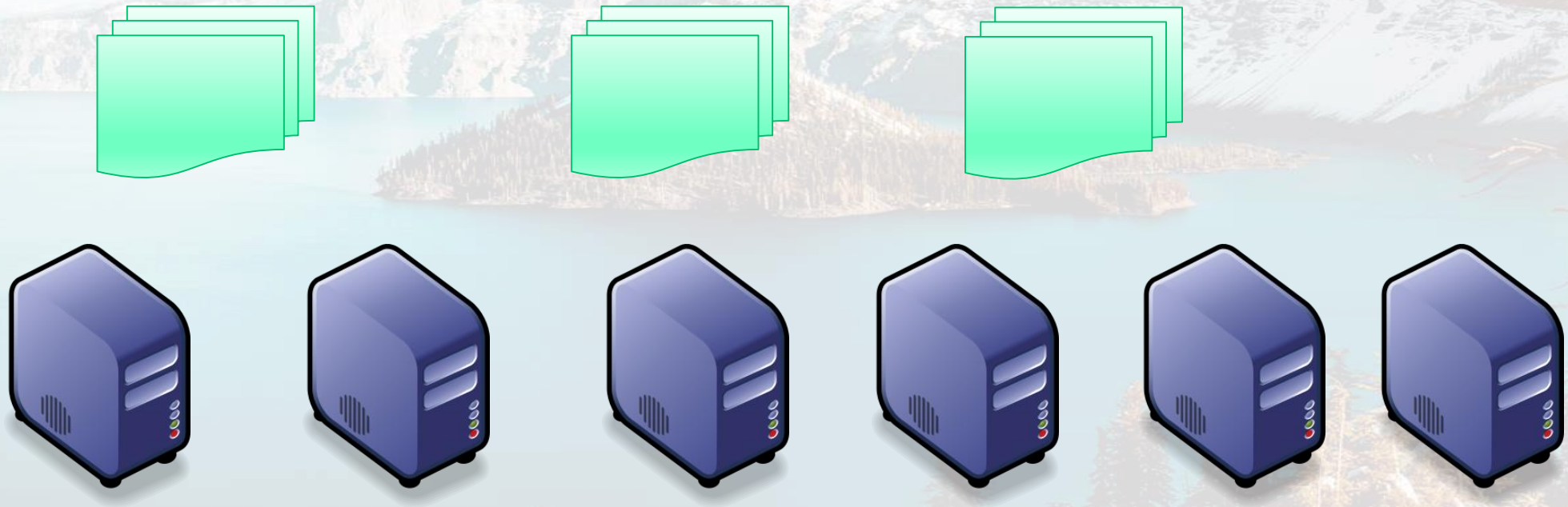
“

Bernard Marr, 2015, "Big Data: 20 Mind-Boggling Facts Everyone Must Read“

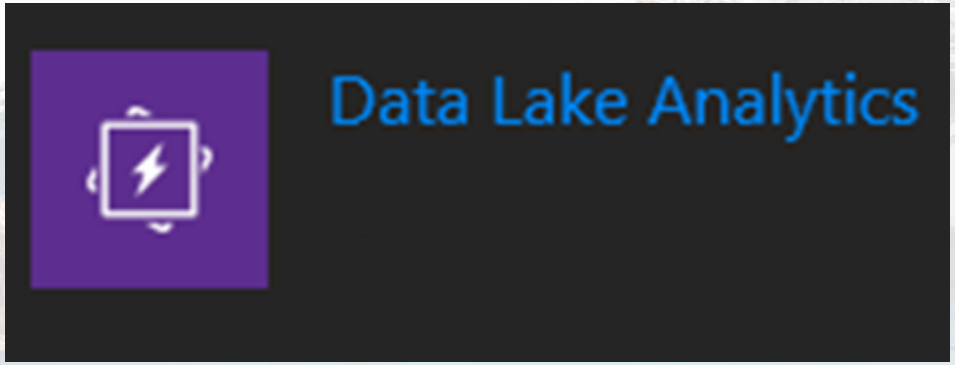
https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/

Data Lake

Storage



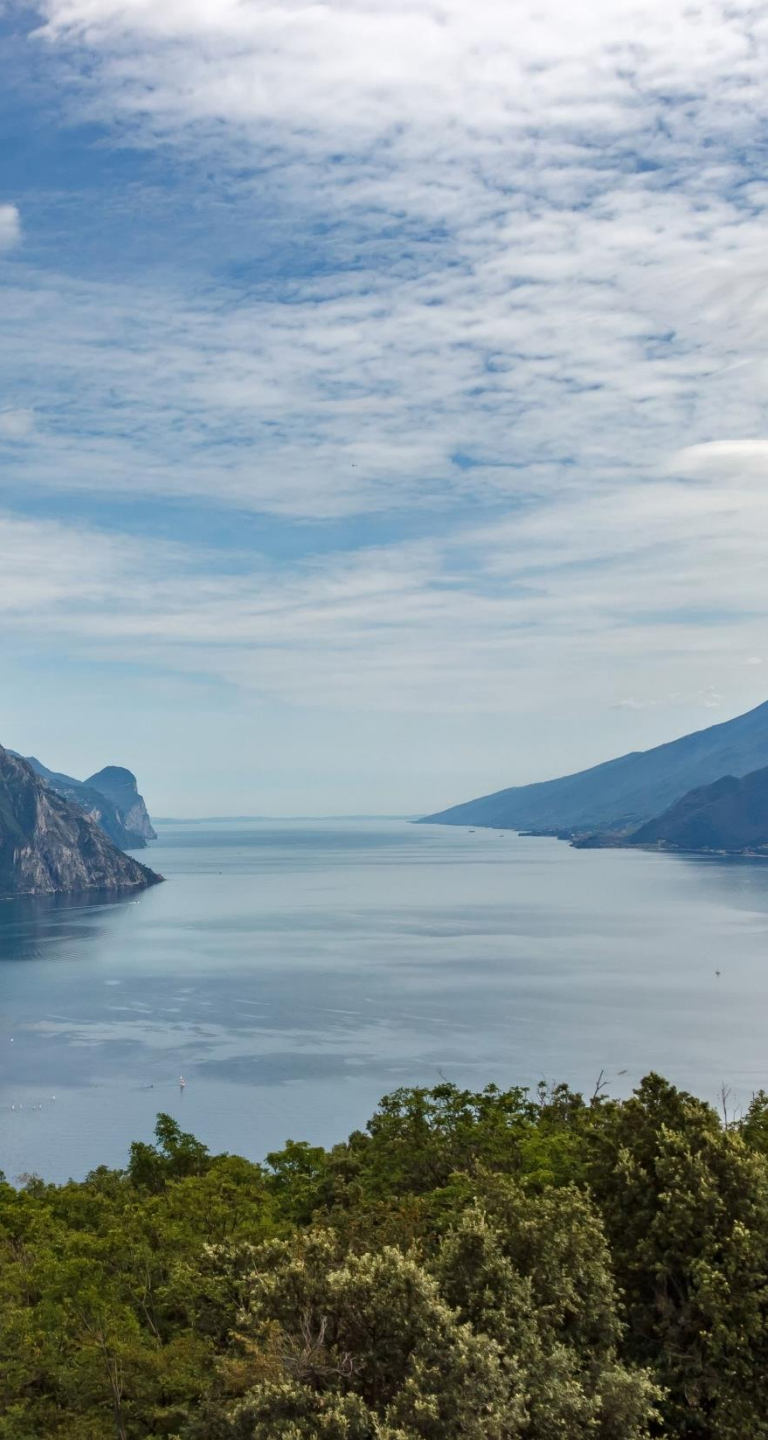
Analysis tools



MapReduce

U-SQL

Target



U-SQL Extractors

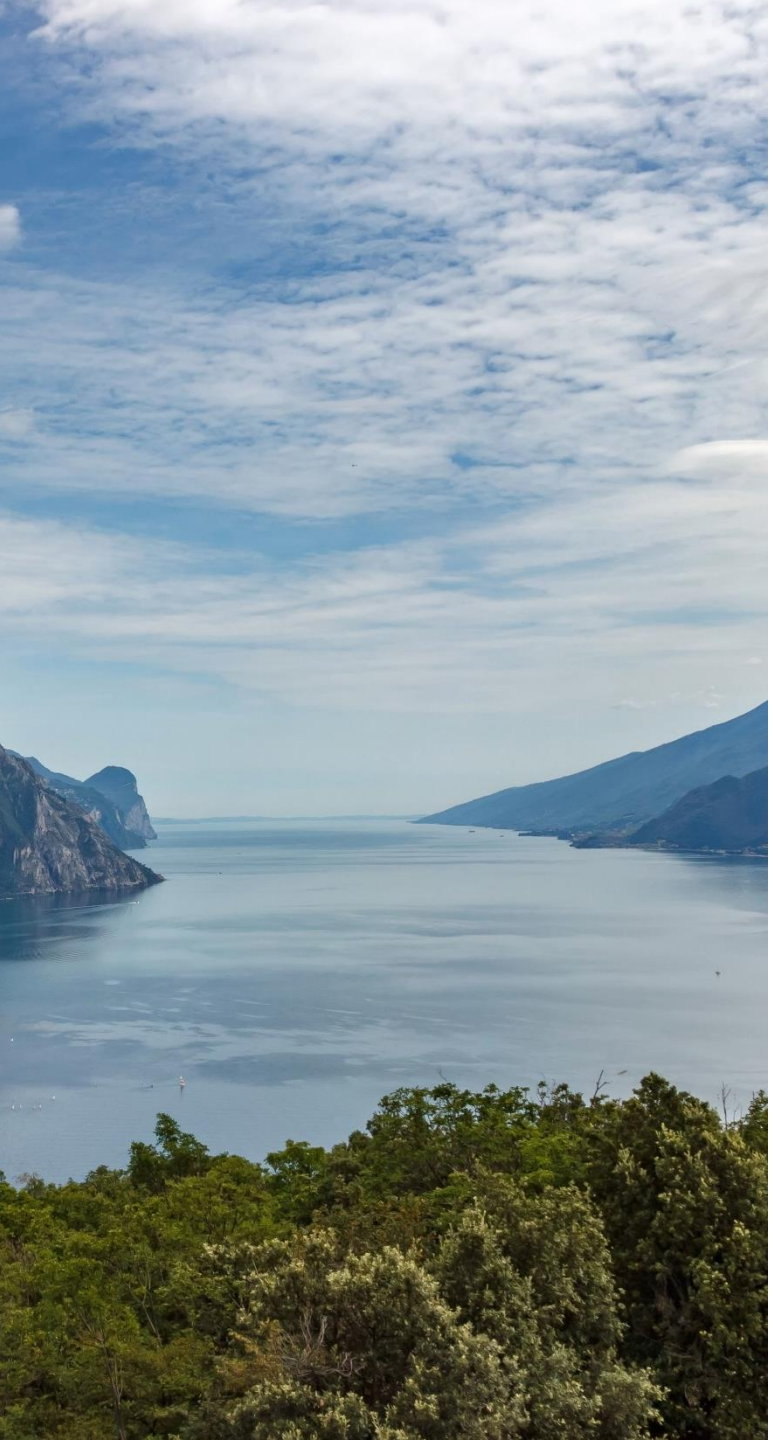
Built-in extractors

Text extractors:

- Extractors.Csv
- Extractors.Tsv
- Extractors.Text

Cognitive Services

https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-cognitive



Custom Extractors

Github

http://usql.io/

https://github.com/Azure/USQL

Tutorial

https://www.taygan.co/blog/2018/01/06/azure-data-lake-series-working-with-json-part-1

Parquet Adapter

https://github.com/elastacloud/parquet-usql

"Apache Parquet is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language."

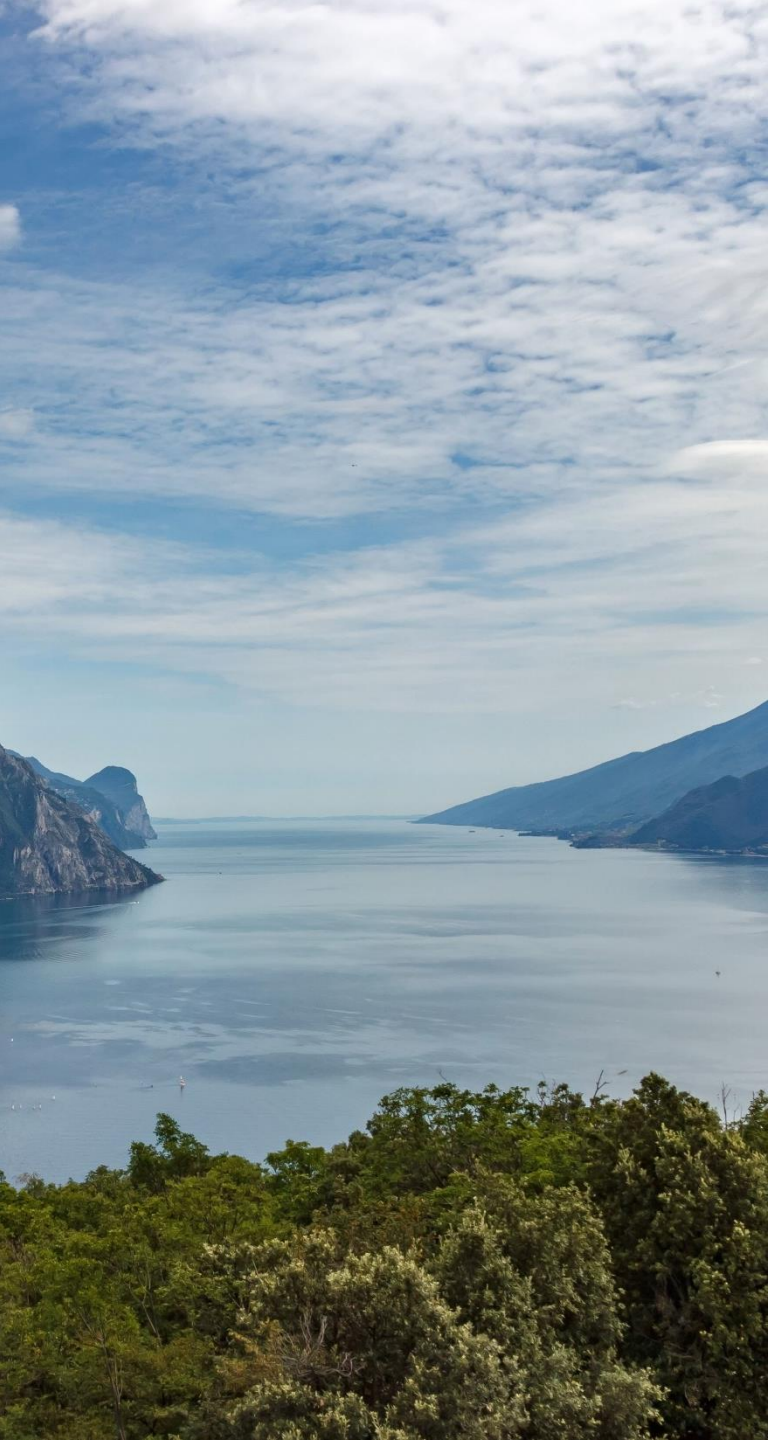
https://parquet.apache.org/

Avro Adapter

Part of the Custom Format Adapter

“Avro is a remote procedure call and data serialization framework developed within Apache's Hadoop project. It uses JSON for defining data types and protocols, and serializes data in a compact binary format. Its primary use is in Apache Hadoop, where it can provide both a serialization format for persistent data, and a wire format for communication between Hadoop nodes, and from client programs to the Hadoop services. “

https://en.wikipedia.org/wiki/Apache_Avro



JSON Extractor

Demo Notes

There are a large number of files in /Weather/WunderGround/history

This file is local

/2016/08/21/

IA_Council_Bluffs_20161107_205556.json

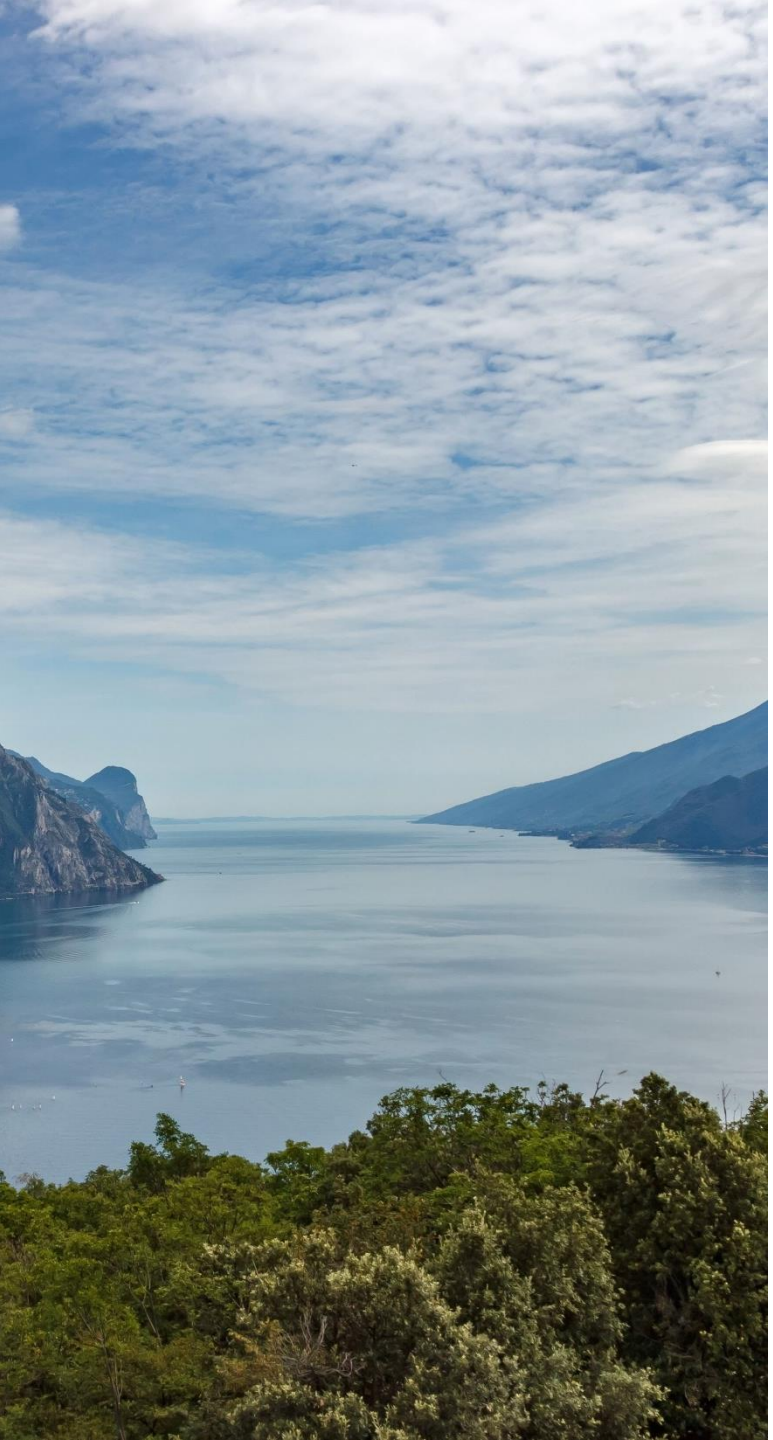
Debugging Assembly

https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-data-lake-tools-local-debug

https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-debug-u-sql-jobs

Excel Adapter

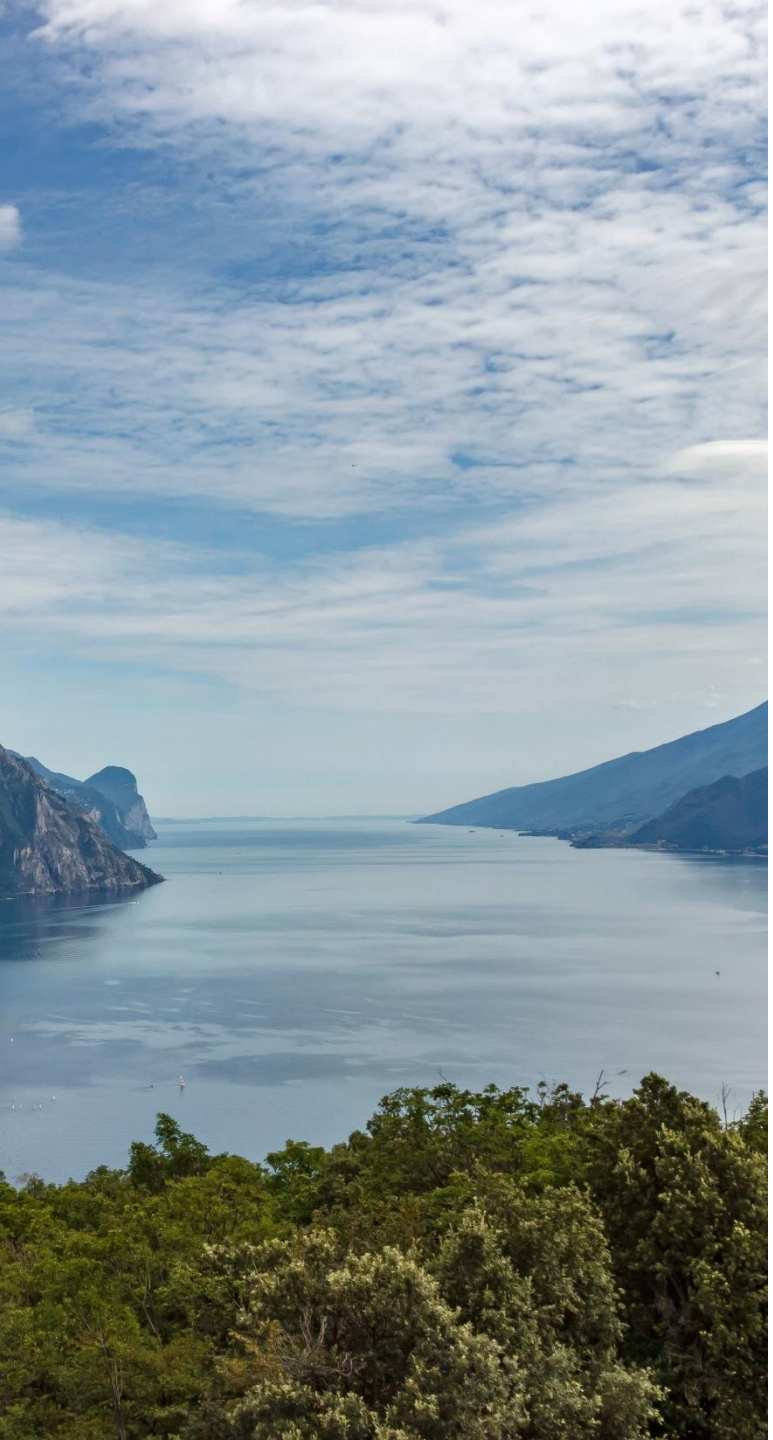
https://github.com/Azure/AzureDataLake/tree/master/Samples/ExcelExtractor



Flexible Schema Extractor

Excel Extractor

https://github.com/Azure/AzureDataLake/tree/master/Samples/ExcelExtractor



Closing

References

http://usql.io/

https://github.com/Azure/USQL

http://www.sqlservercentral.com/stairway/142480/

https://blogs.msdn.microsoft.com/azuredatalake/

https://channel9.msdn.com/Series/AzureDataLake

https://www.census.gov/programs-surveys/acs/data/pums.html

Microsoft Build 2018 Session

BRK4100 Using existing language skillsets to create large-scale, cloud-based analytics

Contact information

RussLoski@SQLMovers.com

www.SQLMovers.com

@sqlmovers

https://www.linkedin.com/in/russloski