

# Michał Smereczyński



Azure Lead Architect @BlueSoft  
Dev<sub>(py)</sub>Ops<sub>(Linux)</sub>

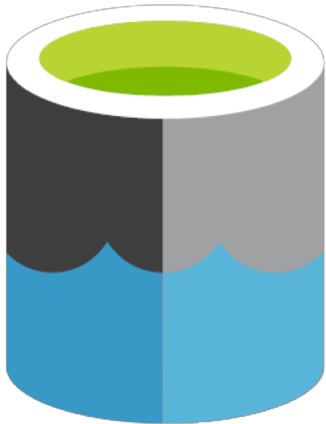


@smereczynski

<https://lnx.azurewebites.net>

<http://AzureTruck.com>

# Data Lake Storage



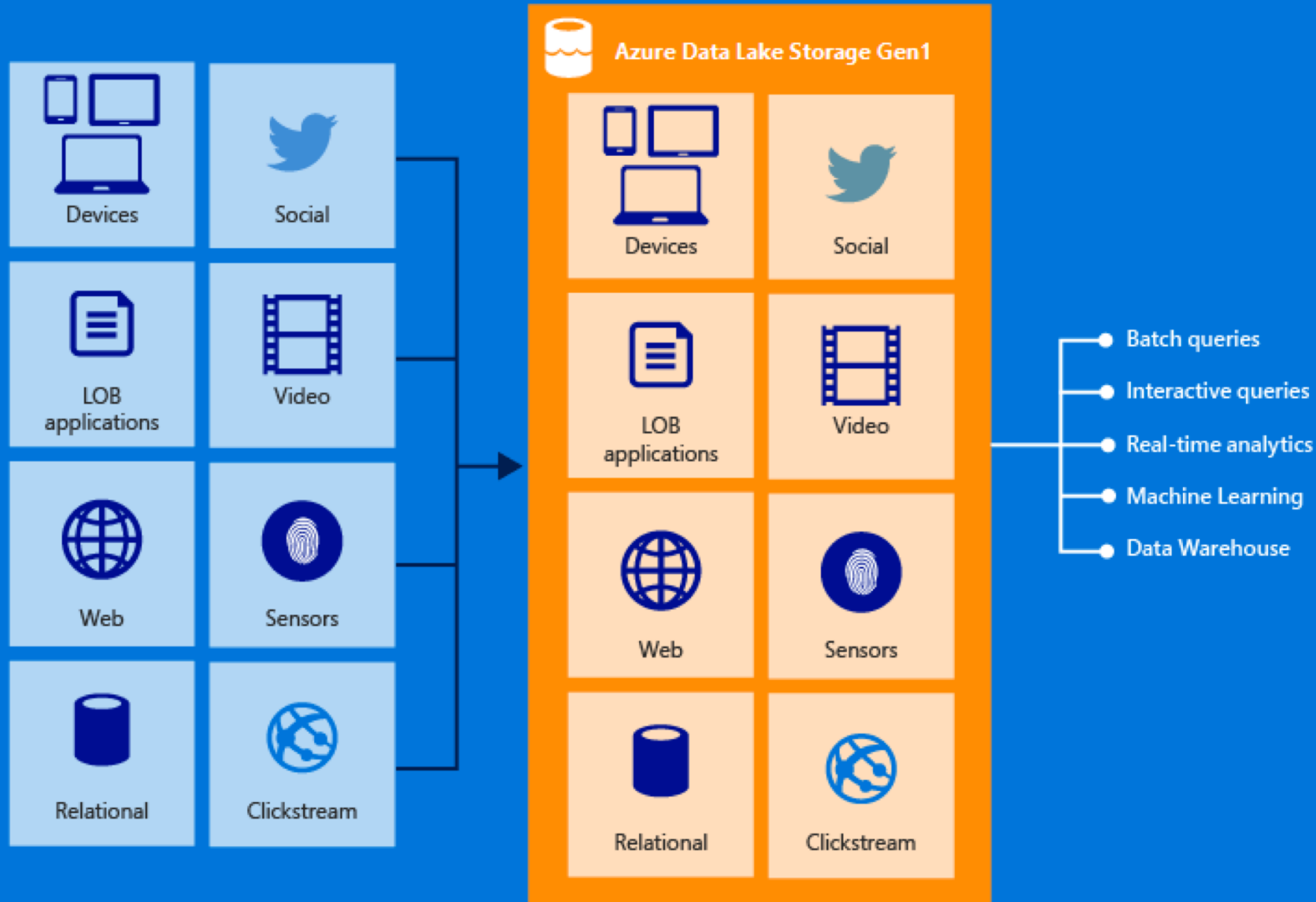
Data Lake Storage Gen1 to usługa oparta o system plików Apache Hadoop kompatybilny z Hadoop Distributed File System (HDFS) i współpracuje z ekosystemem Hadoop.

**Ingest**  
In native format

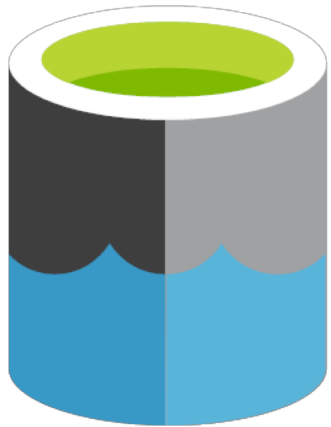
**Prepare**  
Cleanse, enrich,  
annotate, schematize

**Store**  
Retain for present and  
future analysis

**Analyze**  
Using analytic engines  
like hadoop and Spark



# Data Lake Analytics



Azure Data Lake Analytics jest usługą zadań analizy na żądanie, która pozwala uprościć analizowanie danych big data. Można skoncentrować się na pisaniu i uruchamianiu zadań oraz zarządzaniu nimi, a nie na obsłudze rozproszonej infrastruktury.

# Data Lake Analytics

```
@searchlog =  
    EXTRACT UserId      int,  
            Start      DateTime,  
            Region      string,  
            Query        string,  
            Duration     int?,  
            Urls          string,  
            ClickedUrls  string  
    FROM "/Samples/Data/SearchLog.tsv"  
    USING Extractors.Tsv();  
  
OUTPUT @searchlog  
    TO "/output/SearchLog-first-u-sql.csv"  
    USING Outputters.Csv();
```

Be like Bączyk, go hybrid.

# Azure Data Factory



ETL as a Service  
Extract, Transform and Load

