Selecting an Appropriate Storage Service in Microsoft Azure

GETTING TO KNOW DATA

Gary Grudzinskas
CLOUD ENGINEER AND AUTHOR

@garygrudzinskas

Objectives

Know how much data is being produced and where it is coming from

Define what structured, semi-structured, unstructured, and streaming data

Understand what a Data Engineer does

Expanding Data

IOT

App

Media

Smart Phone

Satellite

Web

Log

User

How much data?

IDC and EMC project that the global datasphere will grow to 44 Zettabytes by 2020. By 2025, it will grow to 163 Zettabytes!

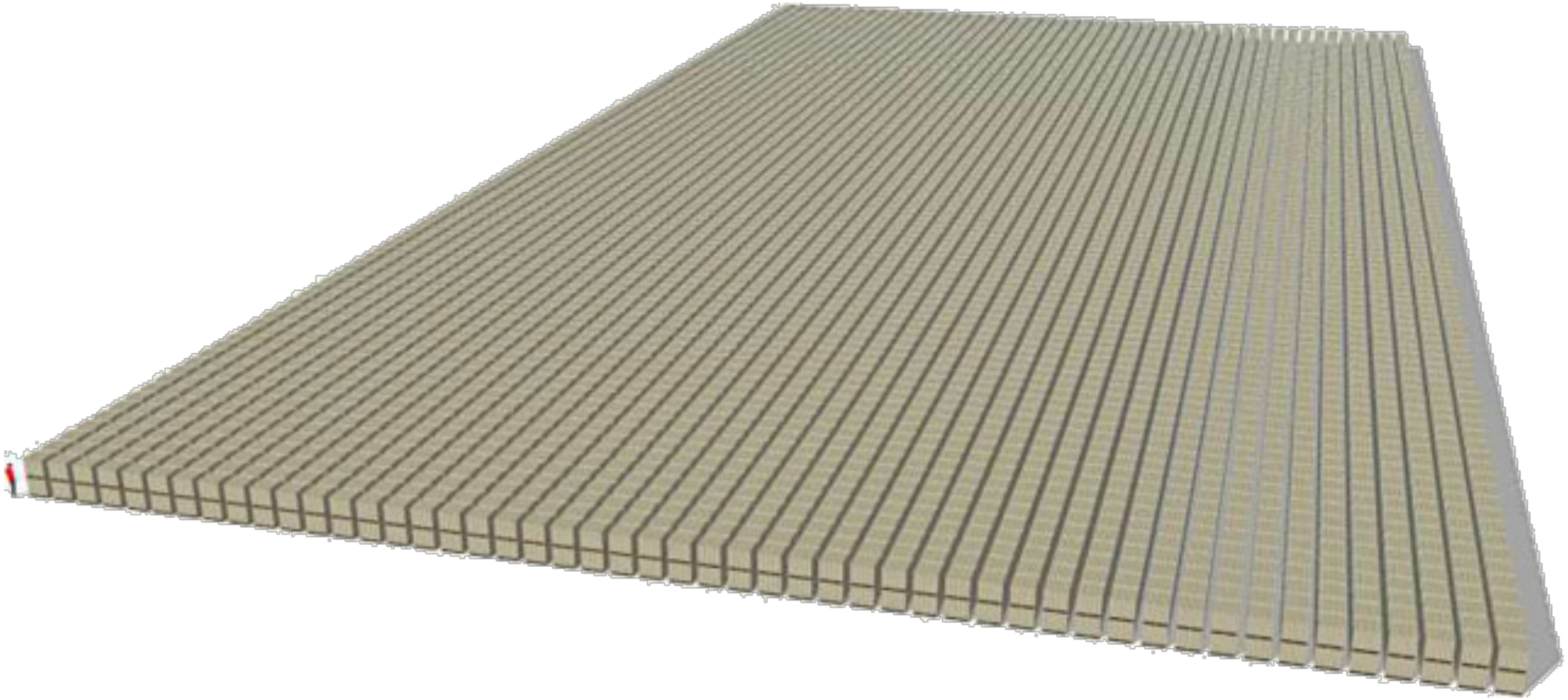
Zettabyte

1,000,000,000,000,000,
000,000,000 Bytes

Zettabyte

1 Trillion Gigabytes

Zettabyte



Binge Watching a Zettabyte?

36 Million Years

Data Engineer

Develop, construct, test and maintain data architectures then integrate, consolidate, and cleanse the data and structure it for use in analytics.

Data Engineer Tasks

Manage and secure the flow of structured, semi-structured, unstructured, and streaming data

Build massive reservoirs for big data

Design, build, and integrate data from various resources, and manage big data

Collaborate with business stakeholders to identify and meet data requirements

Optimize the performance of big data ecosystems

Other Data Roles

Data Scientist

Perform advanced analytics to extract value from data.

Database Administrator

Perform the administration, maintenance, backup, and performance tuning of databases

Data Engineer

Wrangle data for data scientists

Structured Data

Data that is organized and ready to seamlessly integrate into a database. It has a strictly defined schema which defines field names, data types, and the relationship between tables.

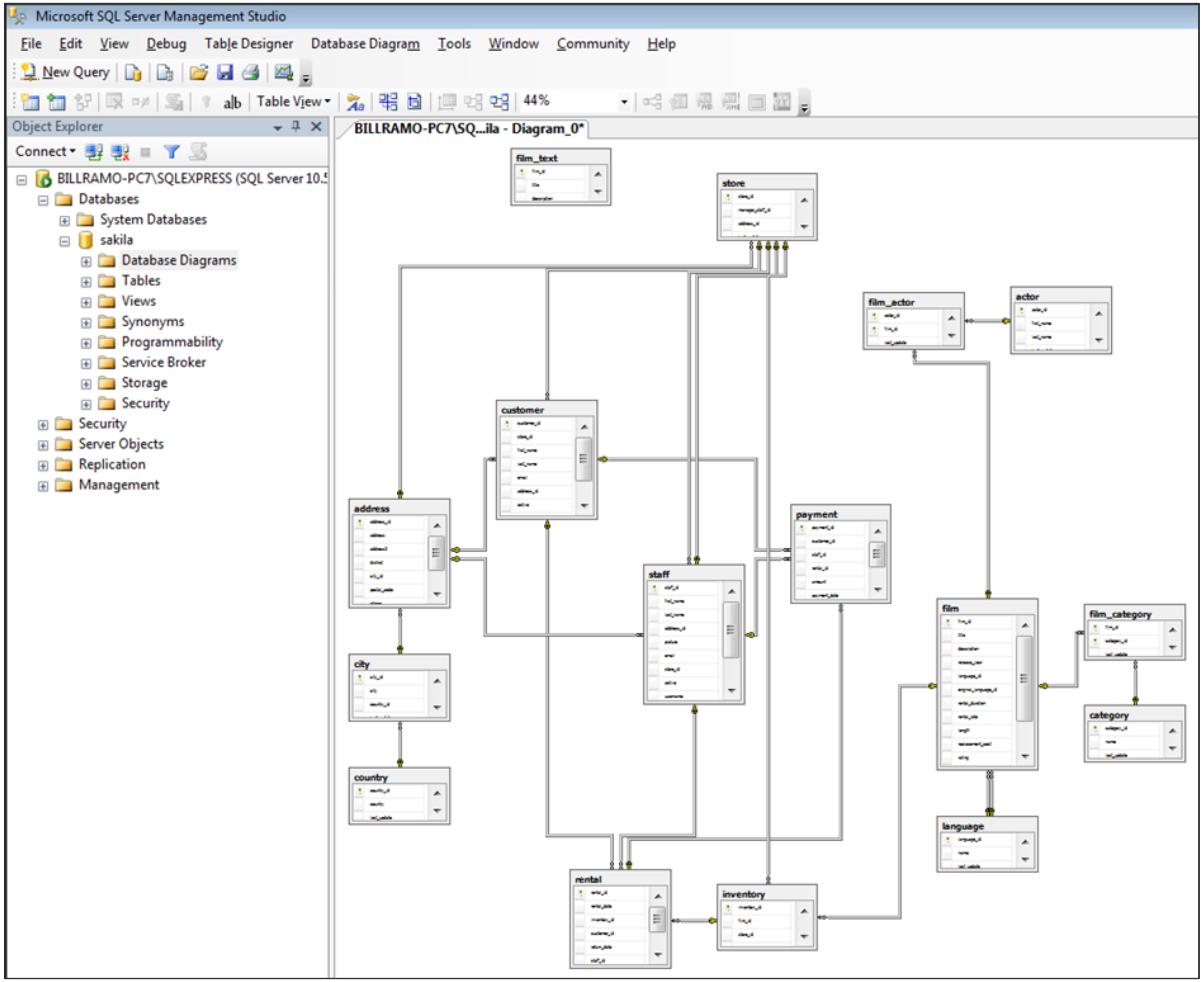
Structured Data Types

SQL DB

Excel

Tables

Key	Address	Phone	City	Gender	Name


Microsoft SQL Server Management Studio
File
Edit
View
Debug
Table Designer
Database Diagram
Tools
Window
Community
Help
New Query
Table View
44%
Object Explorer
Connect
BILLRAMO-PC7\SQLEXPRESS (SQL Server 10.5
Databases
System Databases
sakila
Database Diagrams
Tables
Views
Synonyms
Programmability
Service Broker
Storage
Security
Security
Server Objects
Replication
Management
BILLRAMO-PC7\SQ...ila - Diagram_0*
film_text
store
film_actor
actor
customer
address
payment
staff
film
film_category
city
category
country
language
rental
inventory

Structured Data

Highly precise schema that is defined on Write

Difficult to make changes to the schema to accept new data changes

Extract Transform Load (ETL)

Semi-structured Data

Data that is not organized and does not conform to a formal structure of tables. But it does have structures such as tags or metadata associated with it. This allows records and fields within the data.

Semi-structured Data Types

JSON

Meta Data

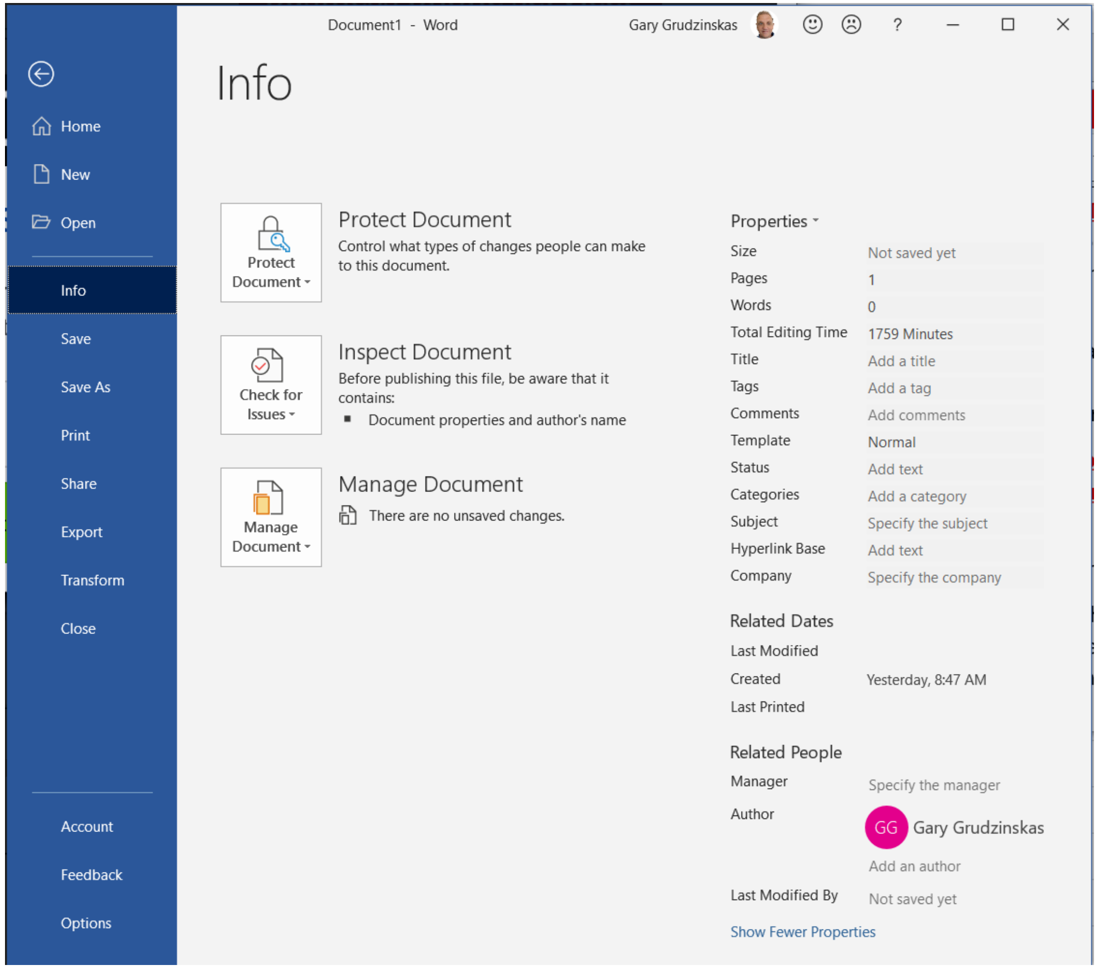
XML

JSON Template

```
{
  "$schema":
"https://schema.management.azure.com
/schemas/2015-01-
01/deploymentTemplate.json#",
  "contentVersion": "",
  "apiProfile": "",
  "parameters": {  },
  "variables": {  },
  "functions": [  ],
  "resources": [  ],
  "outputs": {  }
}
```

```
"parameters": {
  "<parameter-name>" : {
    "type" : "<type-of-parameter-value>",
    "defaultValue": "<default-value-of-
parameter>",
    "allowedValues": [ "<array-of-allowed-
values>" ],
    "minValue": <minimum-value-for-int>,
    "maxValue": <maximum-value-for-int>,
    "minLength": <minimum-length-for-
string-or-array>,
    "maxLength": <maximum-length-for-
string-or-array-parameters>,
    "metadata": {
      "description": "<description-of-the
parameter>"
    }
  }
}
```

Word (or any) Document with Meta Data



Info

Protect Document

Control what types of changes people can make to this document.

Inspect Document

Before publishing this file, be aware that it contains:

- Document properties and author's name

Manage Document

There are no unsaved changes.

Properties

Size	Not saved yet
Pages	1
Words	0
Total Editing Time	1759 Minutes
Title	Add a title
Tags	Add a tag
Comments	Add comments
Template	Normal
Status	Add text
Categories	Add a category
Subject	Specify the subject
Hyperlink Base	Add text
Company	Specify the company

Related Dates

Last Modified	
Created	Yesterday, 8:47 AM
Last Printed	

Related People

Manager	Specify the manager
Author	GG Gary Grudzinskas Add an author
Last Modified By	Not saved yet

Unstructured Data

“Everything else”. Does not have a pre-defined data model and it is not organized in any particular manner that allows traditional analysis.

Unstructured Data Types

Texting

Log

Music

Message

Doc

Conversation

Web

Email

App

Video

90% of all new data is unstructured

Unstructured Data

Does not have a schema or attributes within the data

Highly flexible to accept new changes to the data

Vast assortment of data types and growing everyday

Streaming Data

“Data not at rest”. Data that is in continuous flow from one place to another place. This flow of the data provides an opportunity for immediate analysis or consumption.

Streaming Data Sources

Media

Constantly sends a stream of data to clients.
Examples include Netflix, YouTube, smartphones, and fitness watches.

Satellite

Constantly stream information.
Examples include GPS, surveillance imagery, and telecommunications.

IOT

Produce a constant feed of data.
Examples include driverless cars, manufacturing automation, POS systems, and soon, almost every device imaginable.

Streaming Data Analysis

Batch:

After the stream is stored the data is analyzed to look for patterns and relationships

Real-time:

The data is analyzed during gathering to make an immediate reaction to a trigger