BIG DATA AND DATA ANALYTICS APPENDIX
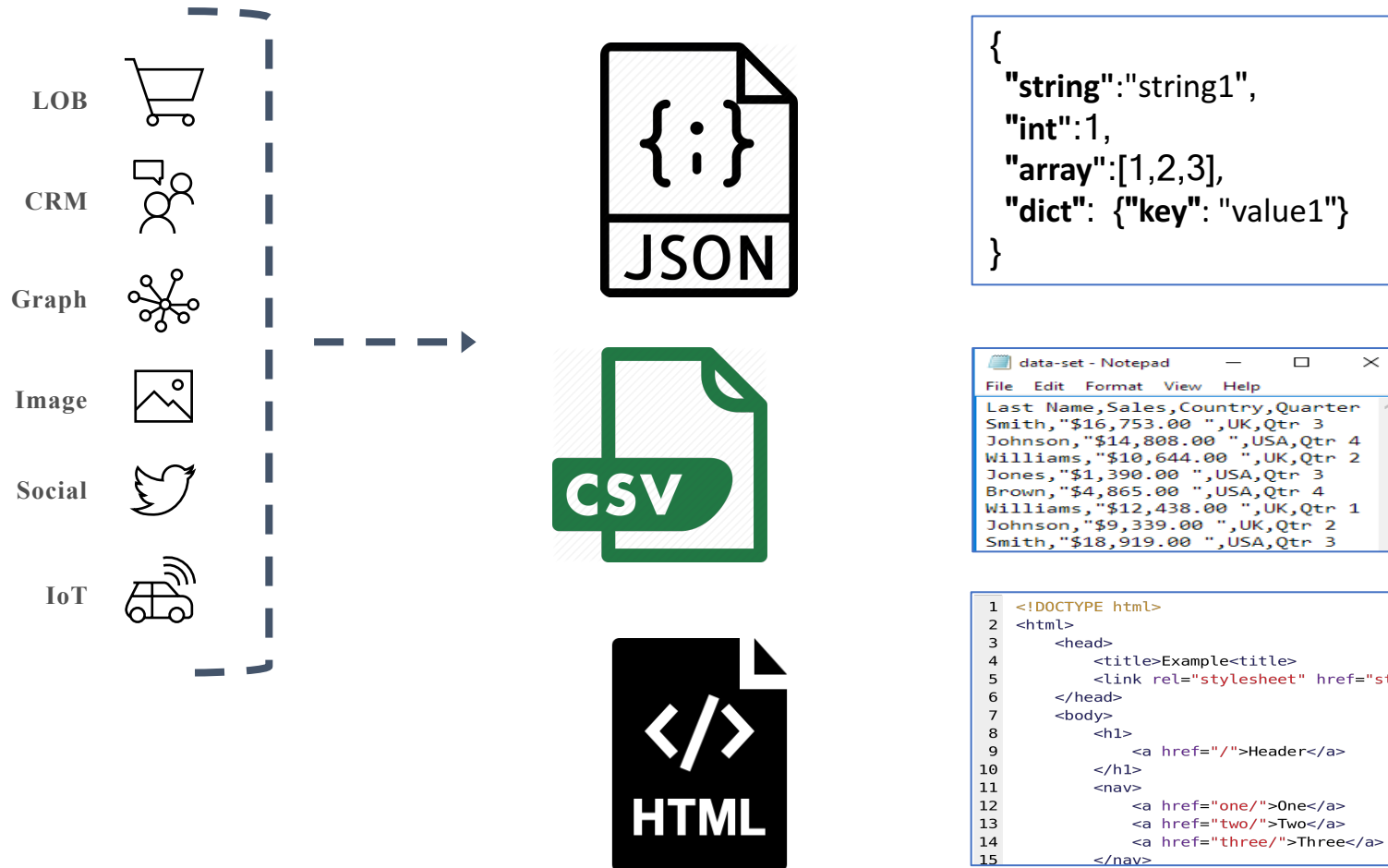
Dr. Anthony Franklin

Big Data File Formats

Common Source Types

LOB

CRM

Graph

Image

Social

IoT

JSON

```
{
  "string":"string1",
  "int":1,
  "array":[1,2,3],
  "dict":  {"key": "value1"}
}
```

CSV

```
data-set - Notepad                    —      □      ×
File   Edit   Format   View   Help
Last Name,Sales,Country,Quarter
Smith,"$16,753.00 ",UK,Qtr 3
Johnson,"$14,808.00 ",USA,Qtr 4
Williams,"$10,644.00 ",UK,Qtr 2
Jones,"$1,390.00 ",USA,Qtr 3
Brown,"$4,865.00 ",USA,Qtr 4
Williams,"$12,438.00 ",UK,Qtr 1
Johnson,"$9,339.00 ",UK,Qtr 2
Smith,"$18,919.00 ",USA,Qtr 3
```

HTML

```
1   <!DOCTYPE html>
2   <html>
3       <head>
4           <title>Example<title>
5           <link rel="stylesheet" href="sty
6       </head>
7       <body>
8           <h1>
9               <a href="/">Header</a>
10          </h1>
11          <nav>
12              <a href="one/">One</a>
13              <a href="two/">Two</a>
14              <a href="three/">Three</a>
15          </nav>
```

Challenges:
- Human Readable
- Not stored in parallel
- Likely inefficient

What file formats are most common when **storing** Big Data?

Data Storage

Optimized Row Columnar

Columnar Store Object:
- Machine readable binary formats (look like gibberish to humans).
- Can be parallelized
- Self-described = carry data schema in file itself.
- Best for read-heavy analytical workloads.

Row-based Store:
- Best for write-heavy transactional workloads

Common Big Data Storage Formats

BIG DATA FORMATS COMPARISON

	Avro	Parquet	ORC
Schema Evolution Support			
Compression			
Splitability			
Most Compatible Platforms	Kafka, Druid	Impala, Arrow Drill, Spark	Hive, Presto
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read

Source: Nexla analysis, April 2018