Büyük Veri

Kısa Giriş

Dr. Cevat Şener

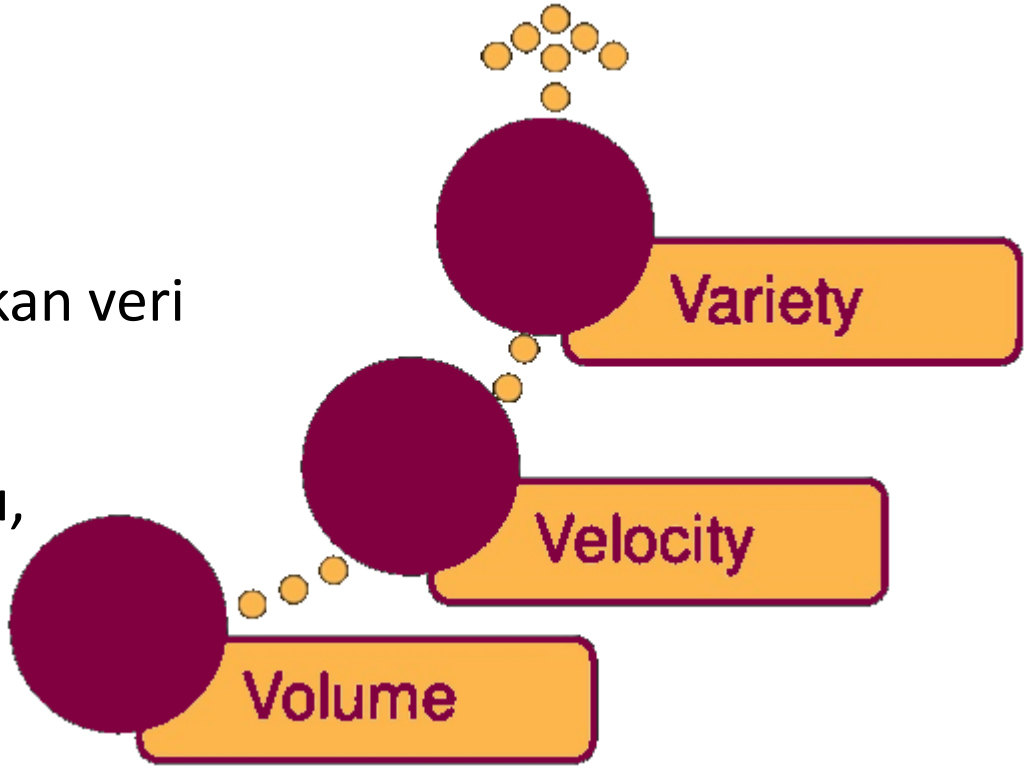
Büyük Veri

- **Klasik sistemlerle veya araçlarla ele alınmayacak ölçüde**
  - **büyük,**

    TB → PB → EB
  - **hızlı üretilen**

    Yüzlerce sensörden akan veri
  - **ve çeşitli…**

    Facebook paylaşımları,
    Sensör verileri,
    Videolar…



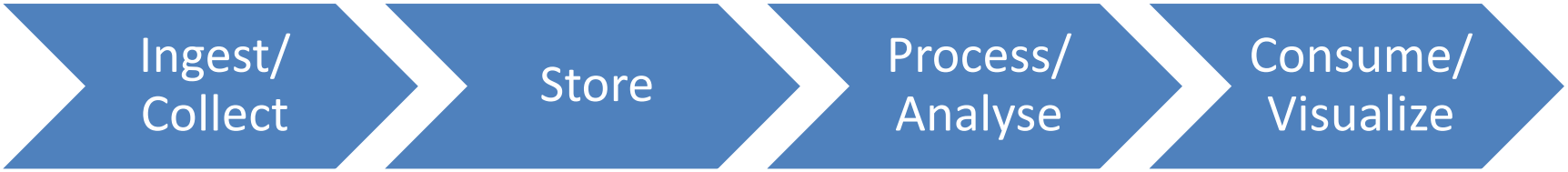
Attributes of Big Data
"V-words"

Volume	quantity of data generated
Velocity	speed at which data is generated and it moves around
Variety	different forms of data
Variability	data with changing meaning
Veracity	fidelity (or quality) of data, which is mainly deformed by inconsistency, incompleteness, biases, noise and abnormality in data collected
Validity	correctness and accuracy of data for the intended use
Volatility	how long data is valid, and hence how long it should be stored
Visibility	visualization capability of data – either directly (an image or a chemical structure) or indirectly (through statistical graphical tools)
Viscosity	resistance to flow in the volume of data where this resistance can come from different data sources, friction from integration flow rates, and processing required to turn the data into insight
Virality	how quickly data is spread and shared to each unique node
Value	importance of data established by the valuable information achieved through effective data mining and analytics on data
…	…

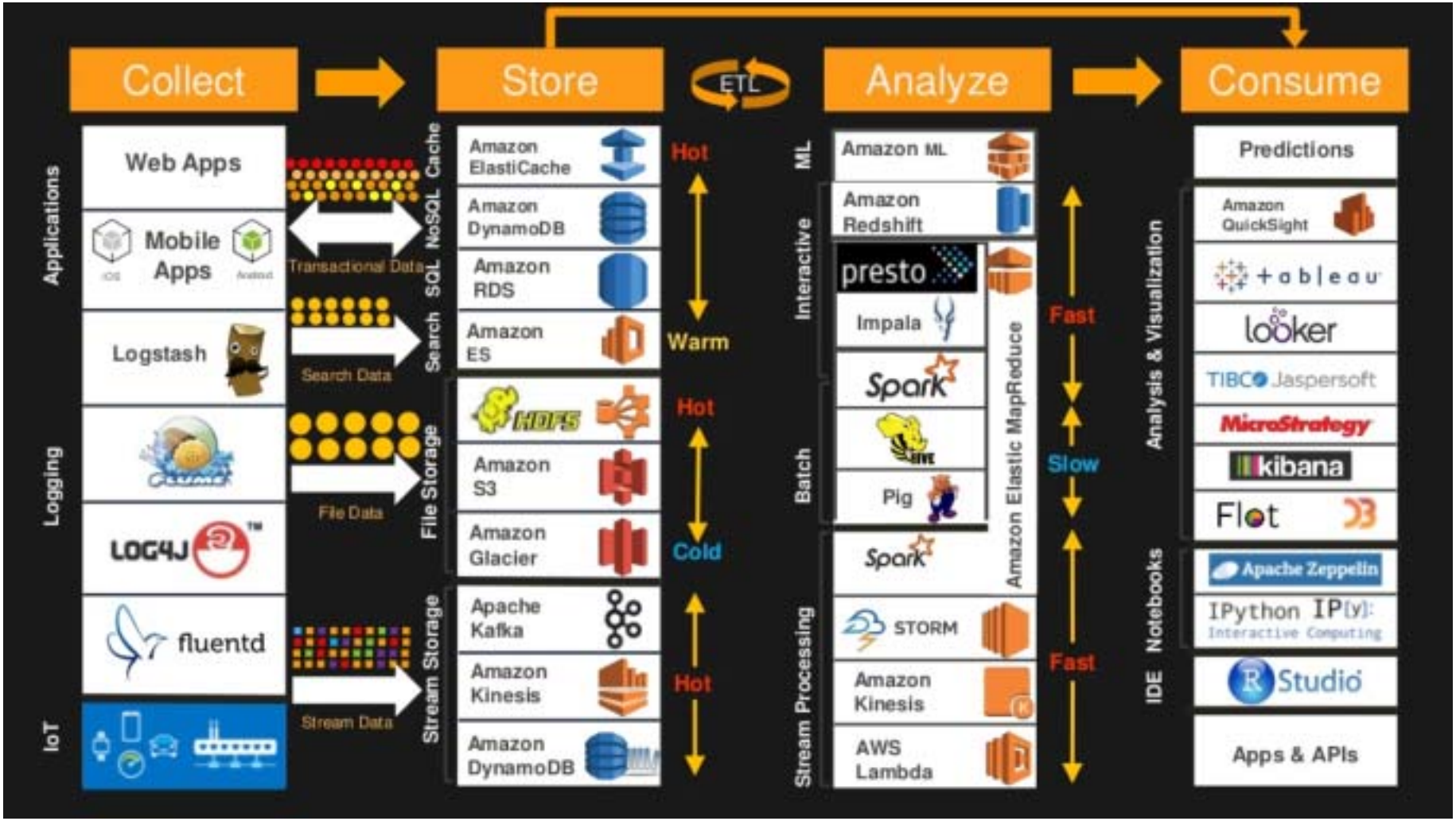
Walmart örneği

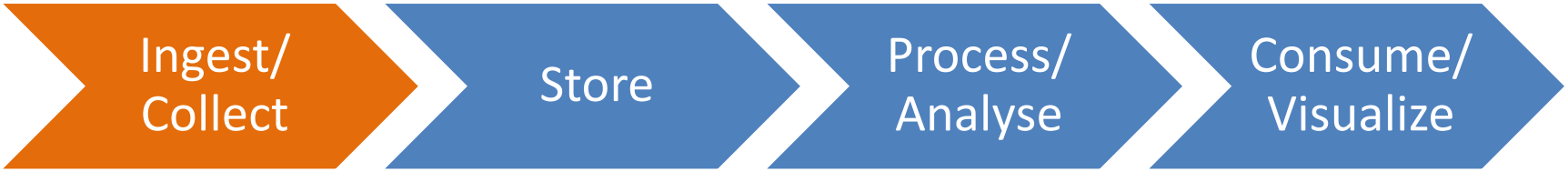
- 245M müşterinin 11K mağaza ve 10 web sitesi ziyareti
- Sosyal medyada haftada 300K kez anılma
- Saatte 1M müşteri işlemi
- Toplam 2,5+ PB yapılandırılmamış veri
  - Alış veriş hareketleri, Sosyal medya verisi, Sosyal medyada öne çıkanlar, Lokal aktiviteler, Hava durumu, …
- Hadoop tabanlı Polaris platformunda yapılan analizler ile %10-15 (> yılda milyar $) gelir artışı

Big Data Processing
Major Phases

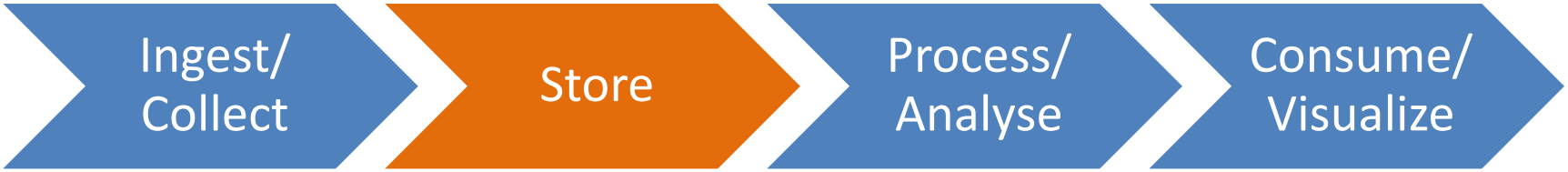


Big Data Processing
Tools





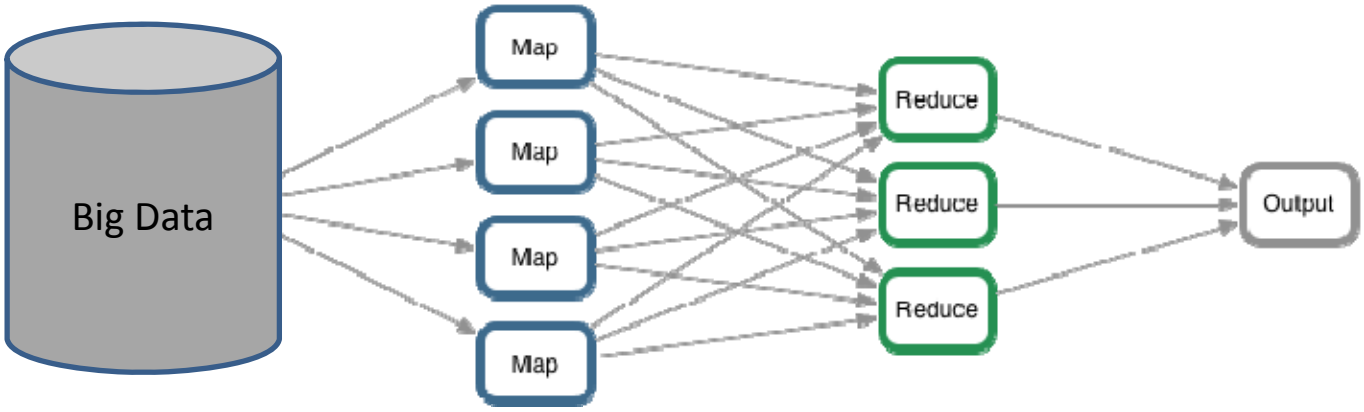
- Kafka
- Kinesis
- Flume
- Sqoop
- …



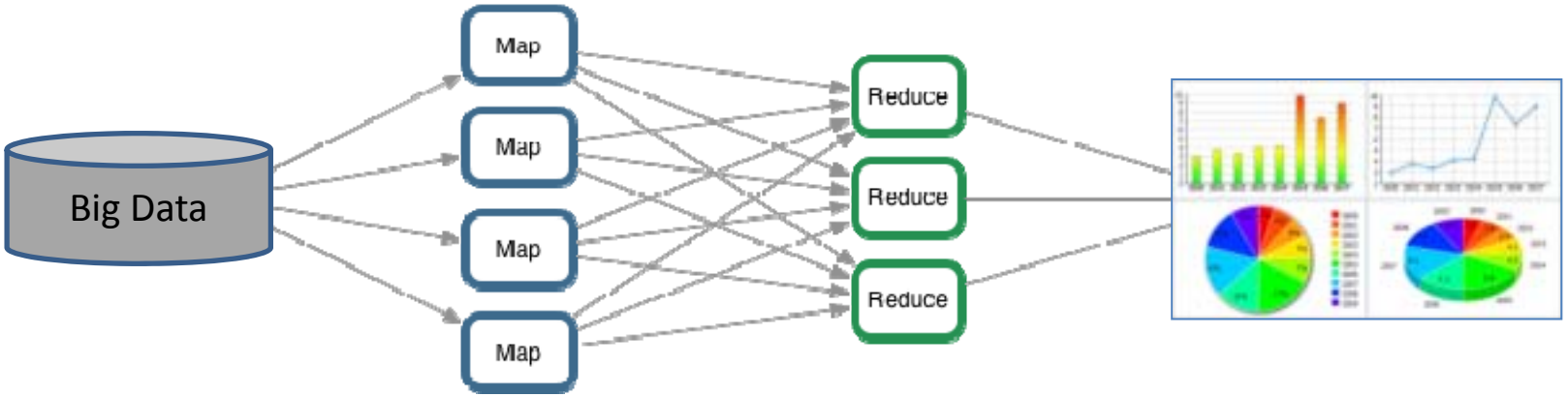
- NoSQL: MongoDB, Cassandra, HBase
- NewSQL: MySQL Cluster, Google Spanner, NuoDB
- InMemoryDB, Cache: Redis, Hazelcast
- File System: HDFS
- …

- Batch Analytics/Processing
- Interactive Analytics
- Real-time Analytics / Stream Processing

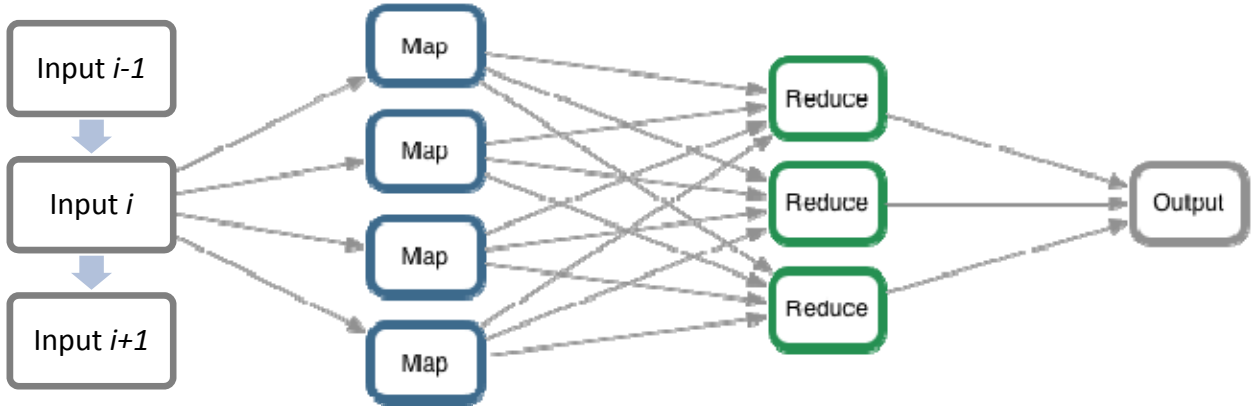
- Batch Analytics/Processing
  - On bulk data
  - Minutes or hours to get answer back
  - Example: Regular (daily, weekly, monthly, …) reports
  - Tools: MapReduce, Hive, Pig, Spark



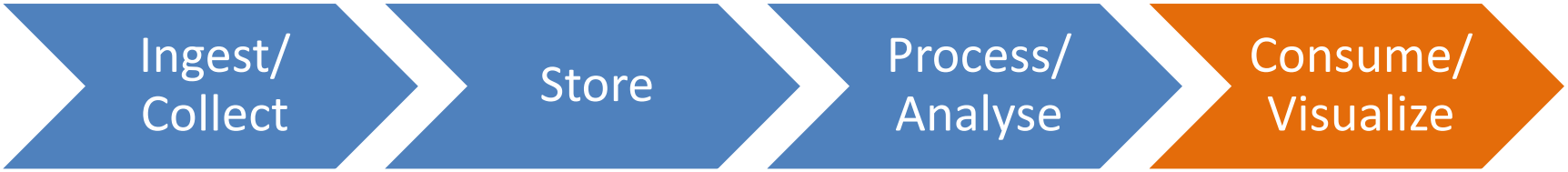
- Interactive Analytics
  - On large amount of data
  - Seconds to get answer
  - Example: Self-service dashboards
  - Tools: Redshift, Presto, Impala, Spark



- Real-time Analytics / Stream Processing
  - On small amount of “hot” data
  - Milliseconds/seconds to get answer back
  - Examples: Event monitoring such as Billing/Fraud alerts, Micro-batch analytics like Calculating current HVAC settings
  - Tools: Storm, Lambda, Spark Streaming



- Machine Learning
  - Supervised Learning on Bulk data (Batch) and Classification/Forecasting on Recent data (may involve Batch, Interactive or Stream processing)
  - Unsupervised Learning on Bulk data for Clustering
  - Tools: Mahout, Spark ML



- Visualization for Business and Science: Tableau, Infogram, ChartBlocks, Google Charts, Visual.ly
- Application/Domain-specific solutions for direct consumption of the results

Teşekkürler…