Datasets

client - processes data

data source - container of data

iteration - client asks data source if there are items left and pulls the next one

```
trait Iterator[A] {
  def hasNext: Boolean
  def next(): A
}
```

observation - data source pushes the next available item to the client

```
// Consumer
trait Observer[A] {
  def onNext(a: A): Unit
  def onError(t: Throwable): Unit
  def onComplete(): Unit
}

// Producer
trait Observable[A] {
  def subscribe(obs: Observer[A]): Unit
}
```

Traversal

breadth-first - from a node, visit its neighbours first, then its children

depth-first - from a node, visit its children first, then its neighbours

Operations

Element-wise

- apply a function to each individual message

conversion - convert values of type A to type B

filtering - only present items that match a condition

projection - only present parts of each data item

Aggregations

- group multiple events together and apply a reduction

left reduction/folding - traverse items first to last

right reduction/folding - traverse items last to first

counting elements

distinct elements

numerical aggregations

    mathematical: min, max, count

    statistical: mean, median, stdev

grouping

KV databases/systems

most common format for distributed databases

key - something that identifies a data record

value - the data record (can be a complex data structure)

groupByKey - group the values for each key into a single sequence

reduceByKey - combine all elements mapped by the same key into one

join - return a sequence containing all parts of elements with matching keys

Immutability

key characteristic of data processing - data is never modified in place, instead, operations are applied to create new versions of the data, without modifying the original version

Copy-On-Write - general technique allowing sharing memory for read-only access that deals with writers by copying the modified resource in a private version

ADT	`collection.mutable`	`collection.immutable`
Array	ArrayBuffer	Vector
List	LinkedList	List
Map	HashMap	HashMap
Set	HashSet	HashSet
Queue	SynchronizedQueue	Queue
Tree	—	TreeSet