CS 591 K1: Data Stream Processing and Analytics

Spring 2020

1/21: Introduction

Vasiliki (Vasia) Kalavri
vkalavri@bu.edu

Course Information

- **Instructor**: Vasiliki Kalavri
    - **Office**: MCS 206
    - **Contact:** vkalavri@bu.edu
- **Course Time & Location**: Tue,Thu 9:30-10:45, MCS B33
    - **Office Hours**: Tue,Thu 11:00-12:30, MCS 206

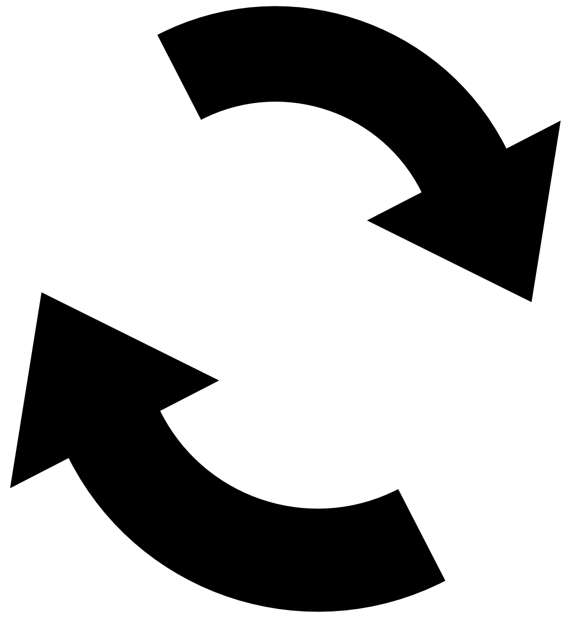
Announcements, updates, discussions

- **Website**: vasia.github.io/dspa20
  - Syllabus: /syllabus.html
  - Class schedule: /lectures.html
    - including today's slides
- **Piazza**: piazza.com/bu/spring2020/cs591k1/home
  - For questions & discussions
- **Blackboard**: learn.bu.edu/...
  - For quizzes, assignment announcements & submissions

What is this course about?

The **design** and **architecture** of modern distributed streaming Systems

- Architecture and design
- Scheduling and load management
- Scalability and elasticity
- Fault-tolerance and guarantees
- State management



- Operator semantics
- Window optimizations
- Filtering, counting, sampling
- Graph streaming algorithms

Fundamental Algorithms for **representing**, **summarizing**, and **analyzing** data streams

Tools

Apache Flink: flink.apache.org

Apache Kafka: kafka.apache.org

Apache Beam: beam.apache.org

Google Cloud Platform: cloud.google.com

Outcomes

At the end of the course, you will hopefully:

- know **when** to use stream processing vs other technology
- be able to comprehensively **compare features** and **processing guarantees** of streaming systems
- be proficient in using Apache Flink and Kafka to build **end-to-end**, **scalable**, and **reliable** streaming **applications**
- have a solid **understanding** of how stream processing systems work and what factors affect their **performance**
- be aware of the **challenges** and **trade-offs** one needs to consider when **designing** and **deploying** streaming applications

Grading Scheme (1)

- **No Exam**
- **5 in-class quizzes (10%)**:
  - Each quiz contributes 2% to the final grade
- **3 hands-on assignments (40%)**:
  - Assignment #1 contributes 10%
  - Assignment #2 contributes 10%
  - Assignment #3 contributes 20%

Grading Scheme (2)

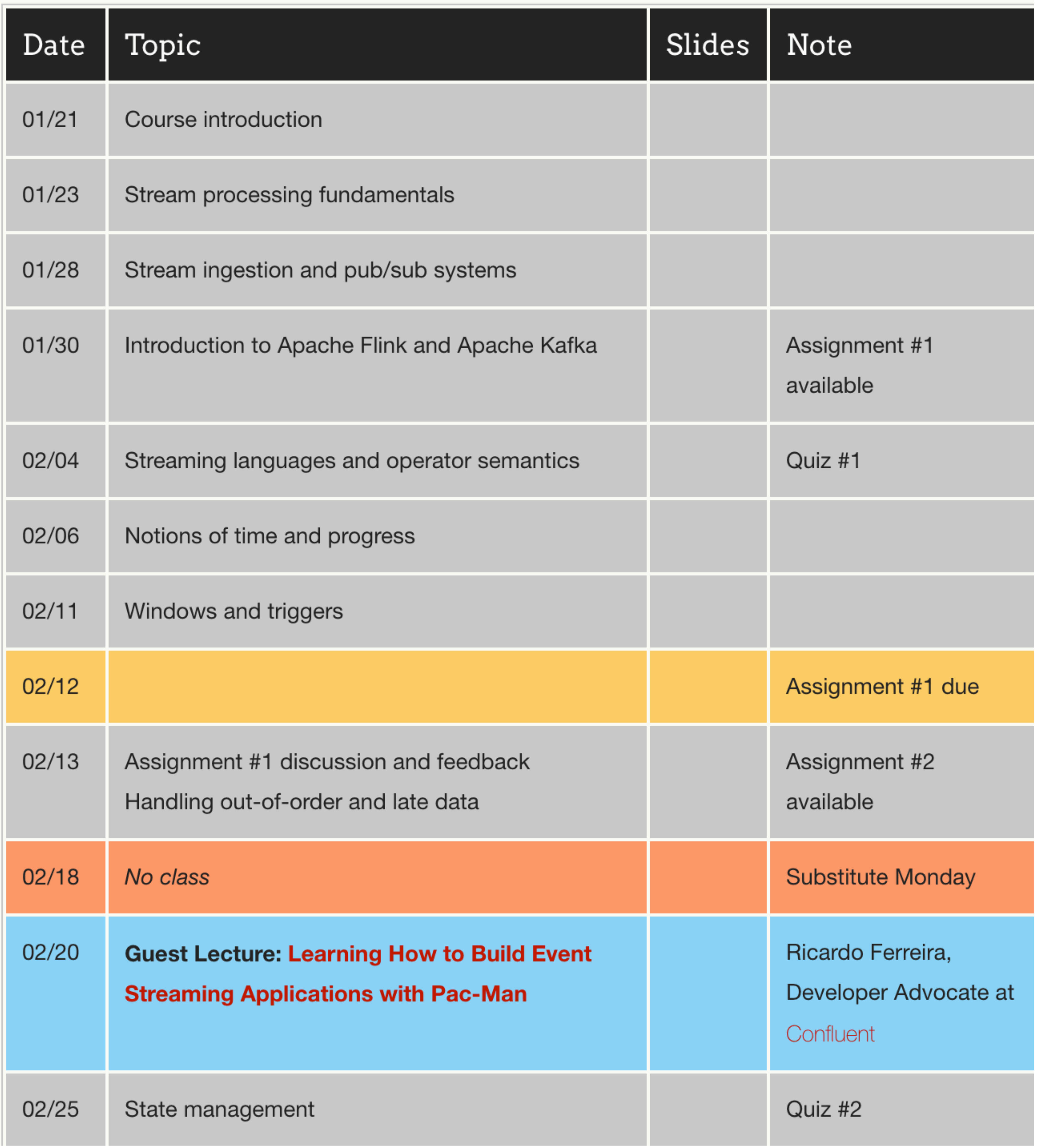
Final Project (50%):

- A real-time monitoring and anomaly detection framework
- To be implemented **individually**

Deliverables

- One (1) **written report** of maximum 5 pages (10%).
- **Code** (including pre-processing, deployment, and testing): (40%)
    - code deliverables must be accompanied by **documentation**

Schedule



Date	Topic	Slides	Note
01/21	Course introduction		
01/23	Stream processing fundamentals		
01/28	Stream ingestion and pub/sub systems		
01/30	Introduction to Apache Flink and Apache Kafka		Assignment #1 available
02/04	Streaming languages and operator semantics		Quiz #1
02/06	Notions of time and progress		
02/11	Windows and triggers		
02/12			Assignment #1 due
02/13	Assignment #1 discussion and feedback Handling out-of-order and late data		Assignment #2 available
02/18	*No class*		Substitute Monday
02/20	**Guest Lecture: Learning How to Build Event Streaming Applications with Pac-Man**		Ricardo Ferreira, Developer Advocate at Confluent
02/25	State management		Quiz #2

vasia.github.io/dspa20/lectures.html

quizzes and announcements

deadline

no class

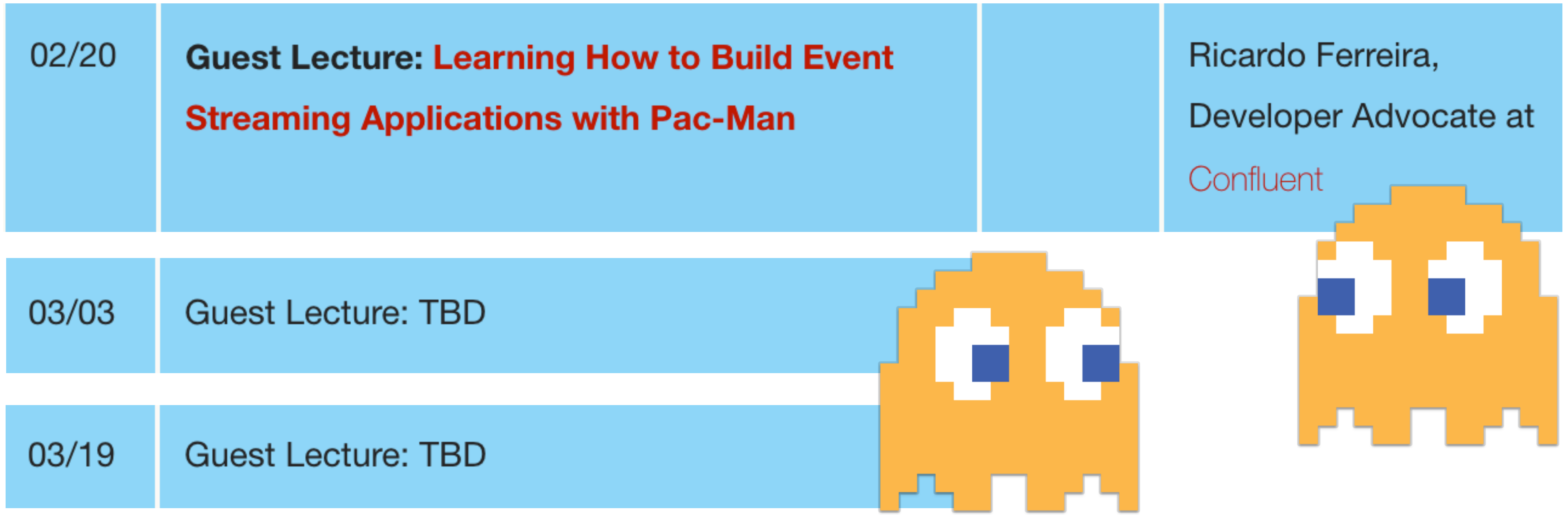
guest lecture

Guest Lectures

- Learn about real-world use-cases of stream processing in industry

- Learn from experts with decades of hands-on experience in building and using distributed systems and data management platforms

- Have fun!



02/20	**Guest Lecture: Learning How to Build Event Streaming Applications with Pac-Man**		Ricardo Ferreira, Developer Advocate at Confluent
03/03	Guest Lecture: TBD		
03/19	Guest Lecture: TBD		

Important dates

Deliverable	Available	Due
Assignment 1	1/30	2/12
Assignment 2	2/13	2/26
Assignment 3	3/3	3/16
Final Project	3/17	4/30

2/18: No Class, Self-study

2/25: Last Day to DROP Clases (without a 'W' grade)

4/3: Last Day to DROP Classes (with a 'W' grade)

Make sure to check the **Official Semester Dates**

Final Project

You will use Apache Flink and Kafka to build a real-time monitoring and **anomaly detection** framework for datacenters.

Your framework will:

- Detect “suspicious” event patterns
- Raise alerts for abnormal system metrics
- Detect invariant violations
- Identify outlier tasks

Interested in a more research-oriented project? Let’s discuss it during office hours.

Inspired by **this paper** : “*SAQL: A Stream-based Query System for Real-Time Abnormal System Behavior Detection*”, USENIX Security '18

Dataset

A subset of traces from a large (12.5k machines) Google cluster

- **https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md**

Make sure to read and become familiar with the format and schema document:

- **https://drive.google.com/file/d/0B5g07T_gRDg9Z0lsSTEtTWtpOW8/view**

Download and play around with “part-00000-of-00500.csv” of:

- job events
- task events
- machine events

Software requirements

- All assignments assume a **UNIX-based** setup.
  - If you are a Windows user, you are advised to use Windows subsystem for Linux (WSL), Cygwin, or a Linux virtual machine to run Flink in a UNIX environment.
- A **Java 8.x** installation. To develop Flink applications and use its DataStream API in Java or Scala you will need a Java JDK. A Java JRE is not sufficient!
- **Apache Maven 3.x**.
- An **IDE** for Java and/or Scala development, such as **IntelliJ IDEA** (preferred), Eclipse, or Netbeans with appropriate plugins installed.
- gsutil for accessing datasets in Google Cloud Storage.

More details: **vasia.github.io/dspa20/exercises.html**

Assignment Submission

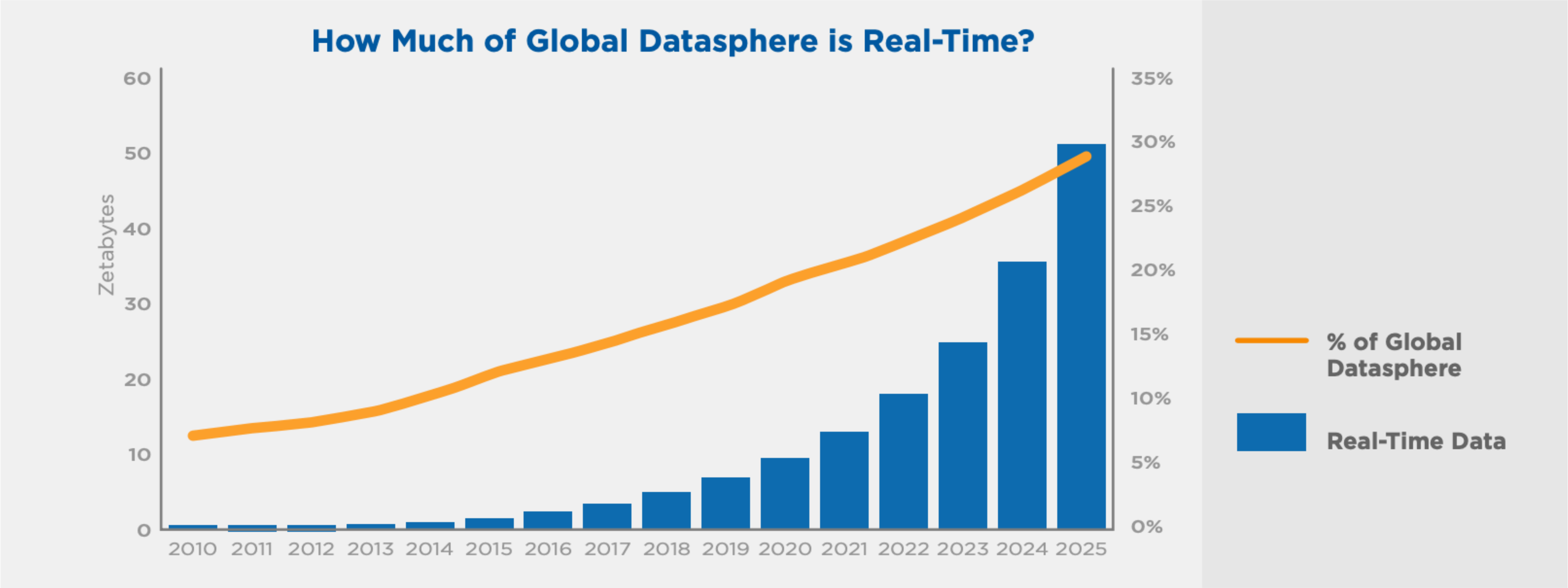
- All assignments and the final project will be submitted via the course **Blackboard**.

- All assignments, as well as the final project, are **due by latest 11:59pm on the day of the respective deadline**.

- *Late* submissions are only eligible for up to **50% of the original score**.

Quiz #0

Why is stream processing important?



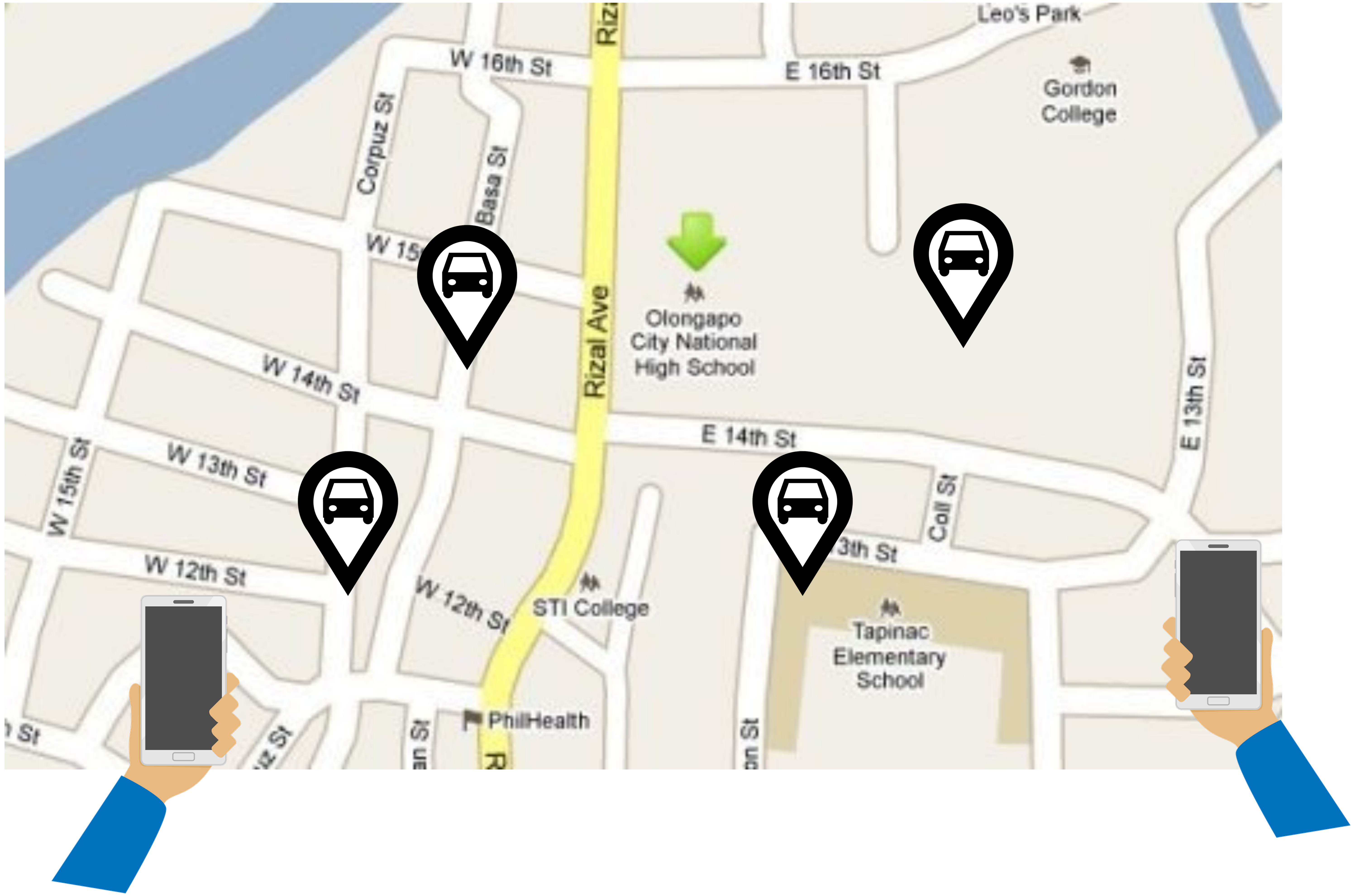
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

By **2025**,
30% of all data
will be **real-time** data.

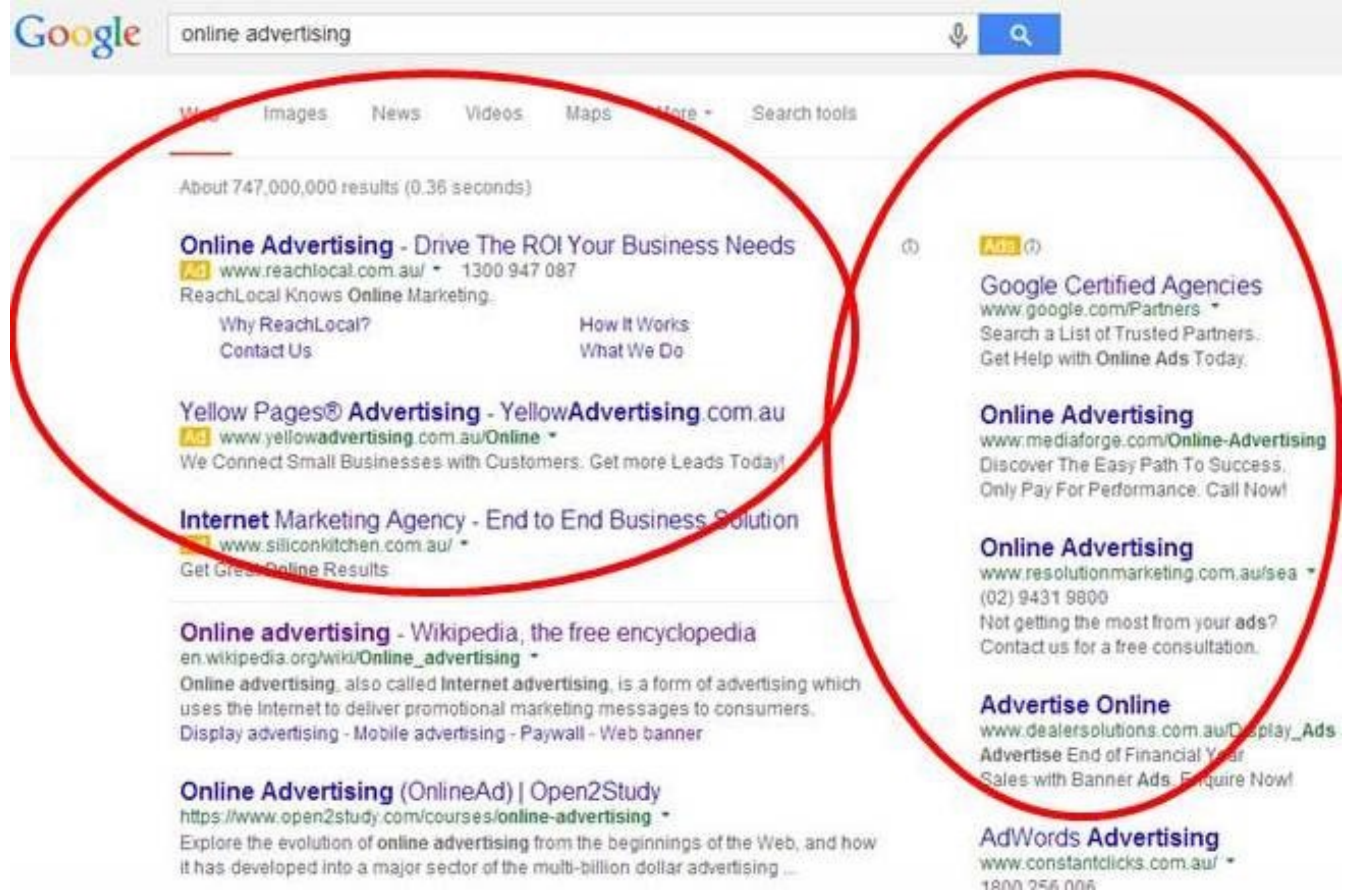
By **2020**,
we will be able to store
less than **15%** of all data.

Can you give me some examples of streaming data sources?

Location-based services



Online recommendations



online advertising

About 747,000,000 results (0.36 seconds)

Online Advertising - Drive The ROI Your Business Needs
Ad www.reachlocal.com.au/ 1300 947 087
ReachLocal Knows **Online** Marketing.
Why ReachLocal? How It Works
Contact Us What We Do

Yellow Pages® **Advertising** - Yellow**Advertising**.com.au
Ad www.yellow**advertising**.com.au/**Online**
We Connect Small Businesses with Customers. Get more Leads Today!

Internet Marketing Agency - End to End Business Solution
Ad www.siliconkitchen.com.au/
Get Great Online Results

Online advertising - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Online_advertising**
Online advertising, also called **Internet advertising**, is a form of advertising which uses the Internet to deliver promotional marketing messages to consumers.
Display advertising - Mobile advertising - Paywall - Web banner

Online Advertising (OnlineAd) | Open2Study
https://www.open2study.com/courses/**online-advertising**
Explore the evolution of **online advertising** from the beginnings of the Web, and how it has developed into a major sector of the multi-billion dollar advertising ...

Ads

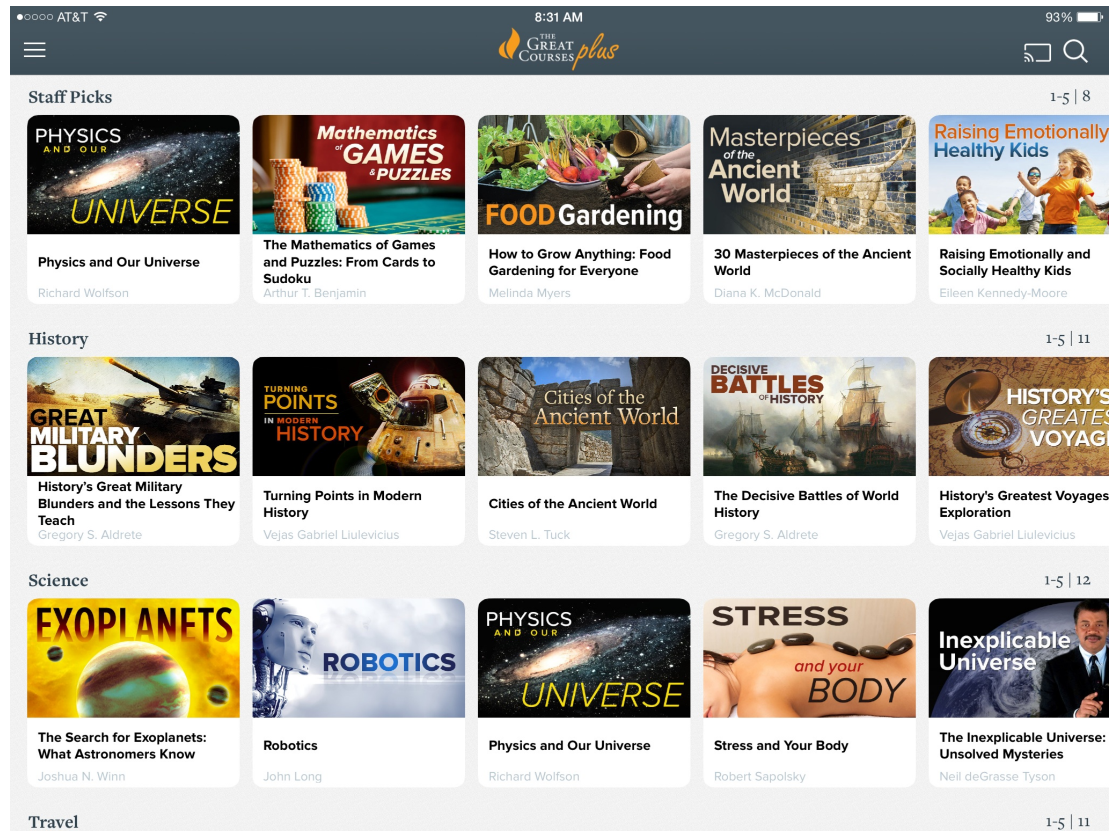
Google Certified Agencies
www.google.com/Partners
Search a List of Trusted Partners.
Get Help with **Online Ads** Today.

Online Advertising
www.mediaforge.com/**Online-Advertising**
Discover The Easy Path To Success.
Only Pay For Performance. Call Now!

Online Advertising
www.resolutionmarketing.com.au/sea
(02) 9431 9800
Not getting the most from your **ads**?
Contact us for a free consultation.

Advertise Online
www.dealersolutions.com.au/Display_Ads
Advertise End of Financial Year
Sales with Banner **Ads**. Enquire Now!

AdWords **Advertising**
www.constantclicks.com.au/
1800 256 006


AT&T
8:31 AM
93%
The Great Courses plus
PHYSICS AND OUR UNIVERSE
Mathematics of GAMES & PUZZLES
FOOD Gardening
Masterpieces of the Ancient World
Raising Emotionally Healthy Kids
GREAT MILITARY BLUNDERS
TURNING POINTS IN MODERN HISTORY
Cities of the Ancient World
DECISIVE BATTLES OF HISTORY
HISTORY'S GREATEST VOYAGES
EXOPLANETS
ROBOTICS
PHYSICS AND OUR UNIVERSE
STRESS and your BODY
Inexplicable Universe

Staff Picks 1-5 | 8

Physics and Our Universe
Richard Wolfson

The Mathematics of Games and Puzzles: From Cards to Sudoku
Arthur T. Benjamin

How to Grow Anything: Food Gardening for Everyone
Melinda Myers

30 Masterpieces of the Ancient World
Diana K. McDonald

Raising Emotionally and Socially Healthy Kids
Eileen Kennedy-Moore

History 1-5 | 11

History's Great Military Blunders and the Lessons They Teach
Gregory S. Aldrete

Turning Points in Modern History
Vejas Gabriel Liulevicius

Cities of the Ancient World
Steven L. Tuck

The Decisive Battles of World History
Gregory S. Aldrete

History's Greatest Voyages Exploration
Vejas Gabriel Liulevicius

Science 1-5 | 12

The Search for Exoplanets: What Astronomers Know
Joshua N. Winn

Robotics
John Long

Physics and Our Universe
Richard Wolfson

Stress and Your Body
Robert Sapolsky

The Inexplicable Universe: Unsolved Mysteries
Neil deGrasse Tyson

Travel 1-5 | 11

Sensor measurements analysis

- Monitoring applications
- Complex filtering and alarm activation
- Aggregation of multiple sensors and joins

- **Examples**
  - Real-time statistics, e.g. weather maps
  - Monitor conditions to adjust resources, e.g. power generation
  - Monitor energy consumption for billing purposes

Stock trading

- Discover correlations, identify trends, forecast future values

Examples

- Find all stocks priced between $20 and $200, where the spread between the high tick and the low tick over the past 30 minutes is greater than 3% of the last price, and where in the last 5 minutes the average volume has surged by more than 300%.
- Find all stocks trading above their 200-day moving average with a market cap greater than $5 Billion that have gained in price today by at least 2%, and are within 2% of today's high.

Financial transaction analysis

- Fraud detection, online risk calculation

Example: Someone steals your phone and sings in your banking app. The app allows transfers of up to €1000 and so the thief makes transfers of €1000 to a "fake account" until either you're out of money or the activity is detected.

- Features to detect fraudulent activity like this:
  - The transaction amount.
  - The number of recent (e.g. the last hour) transactions.
  - Whether money was sent to this recipient account for the first time in the past 24 hours (in other words, to an “unknown” recipient account).

Read more: **https://www.ververica.com/blog/real-time-fraud-detection-ing-bank-apache-flink**

Call monitoring

- Service monitoring, e.g. source and destination phone numbers, their first and last cell towers

Examples:

- Location-based services
- Monitor cell tower load
- Continuously maintain call signatures for fraud detection
  - call frequency
  - top-K cell towers used

Web activity analysis

- Visualization and aggregation
  - impressions, clicks, transactions, likes, comments
- Analytics on user activity
  - Filtering, aggregation, joins with static data (e.g. user profile data)

Examples

- online A/B testing
- trending topics
- sentiment analysis, e.g., reaction to just published campaign
- online recommendations of products, articles, people

Online traffic management

- Analysis of real-time vehicle locations to improve traffic conditions
- Provide real-time scheduling information for public transport
- Optimize transport network flow and recommend alternative routes

Example:

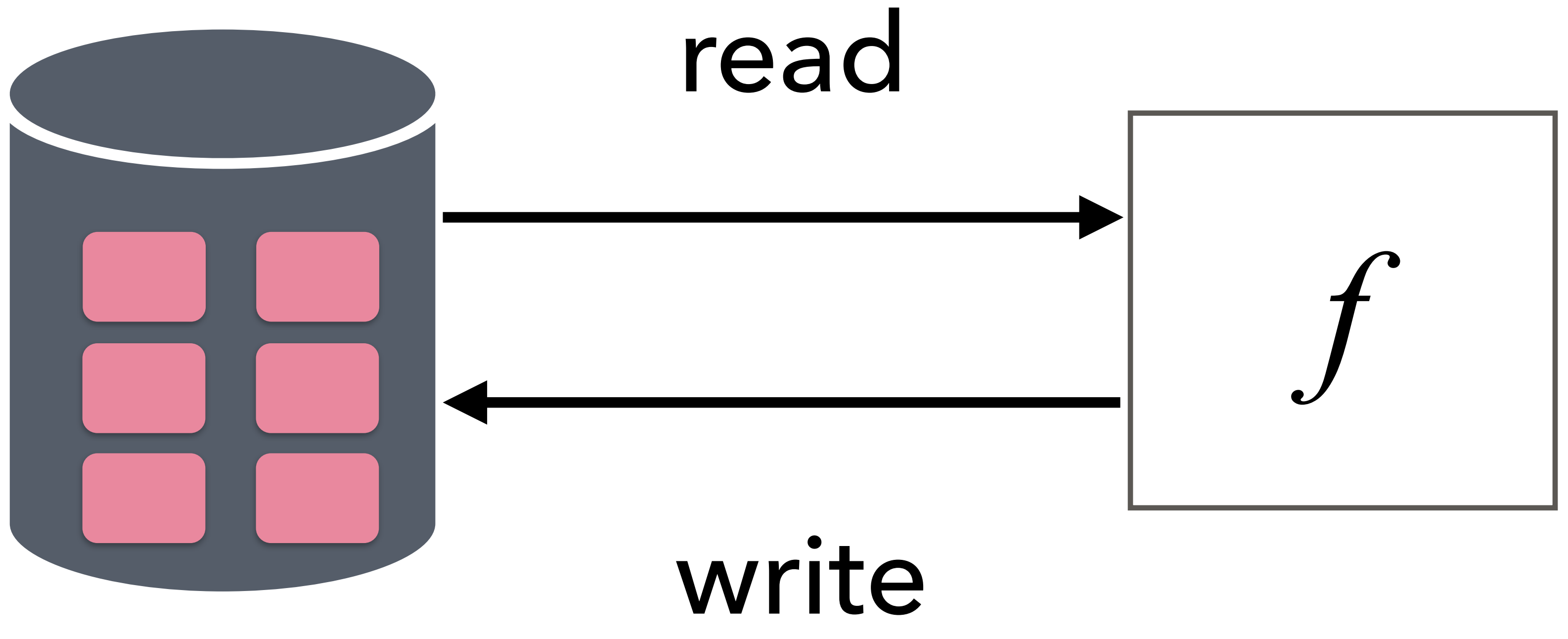
- Alibaba City Brain adjusts traffic lights in real-time to reduce congestion and clear paths for emergency response vehicles
- Read more: **https://edition.cnn.com/2019/01/15/tech/alibaba-city-brain-hangzhou/index.html**

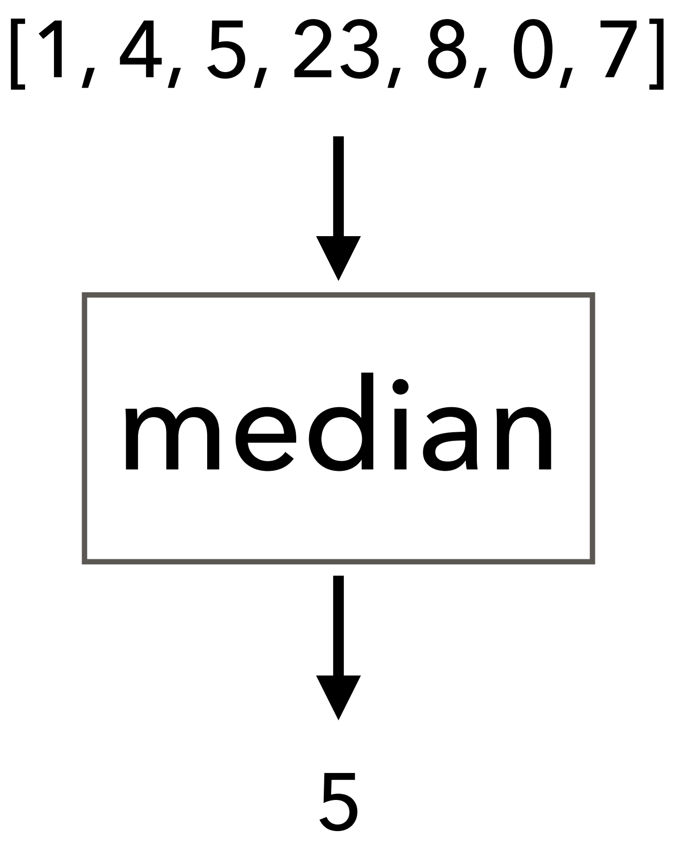
Why is stream processing challenging?

Using pseudocode (or the programming language of your choice), write a program that reads a stream of integers and computes:

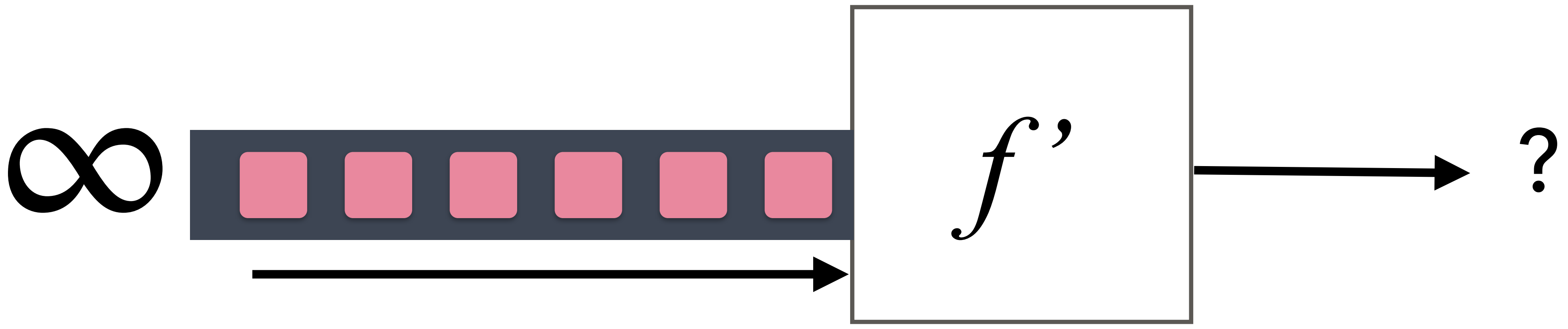
1. the **maximum** number seen so far
2. the **average** of all numbers seen so far
3. the **median** of all numbers seen so far

Complete data accessible in persistent storage





Continuously arriving, possibly unbounded data



- We cannot store the entire stream
- No control over arrival rate or order

Consider a set of 1000 sensors deployed in different locations inside a forest. The sensors monitor temperature and smoke levels and generate a measurement every 5 seconds.

Write a program that every 1 minute emits the average temperature over the last 10 minutes.

Some hard problems in stream processing

Time

Reconfiguration & updates

Debugging

Processing guarantees

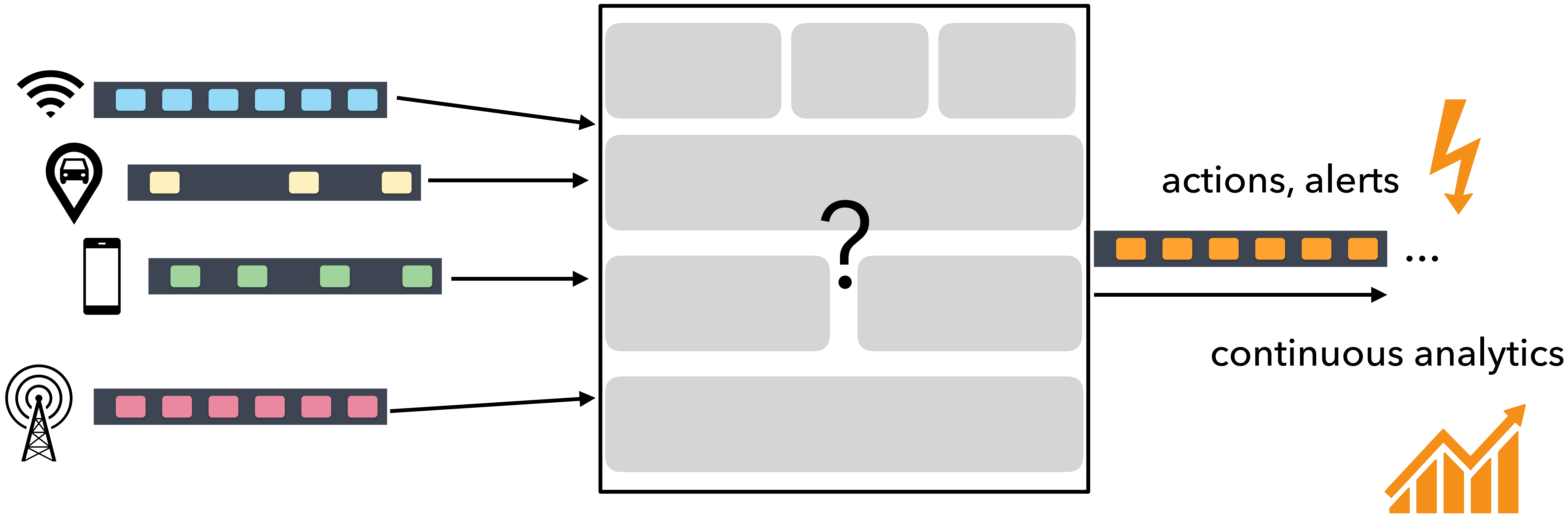
Order

Retractions & results amendment

Progress

Fault-tolerance & high-availability

Building a stream processor…



Optional reading

- The 8 Requirements of Real-Time Stream Processing
  http://cs.brown.edu/~ugur/8rulesSigRec.pdf

- Streaming 101: The world beyond batch
  www.oreilly.com/ideas/the-world-beyond-batch-streaming-101