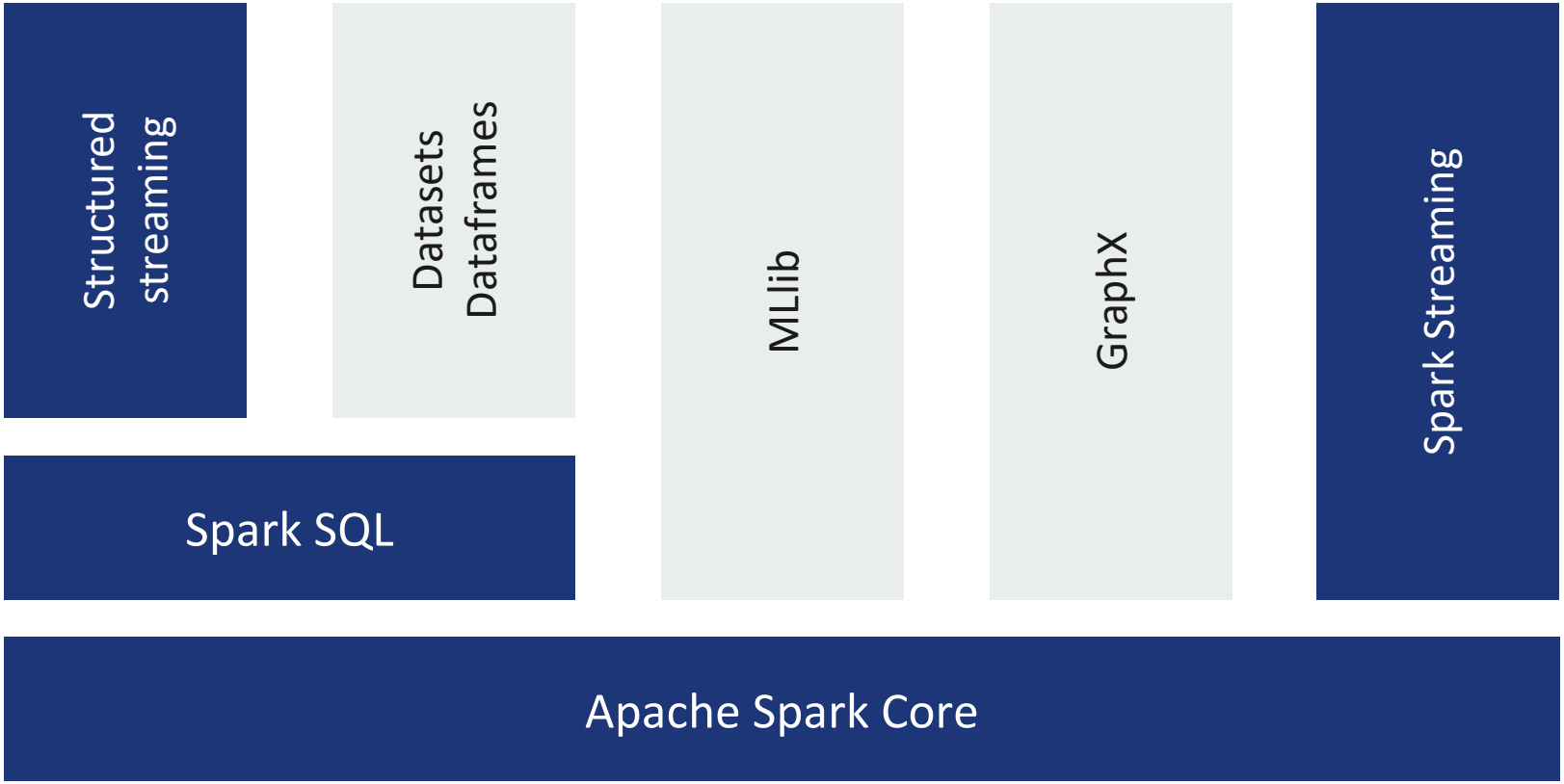
Spark Streaming API

Adrien Gicquel
Hugo Friant

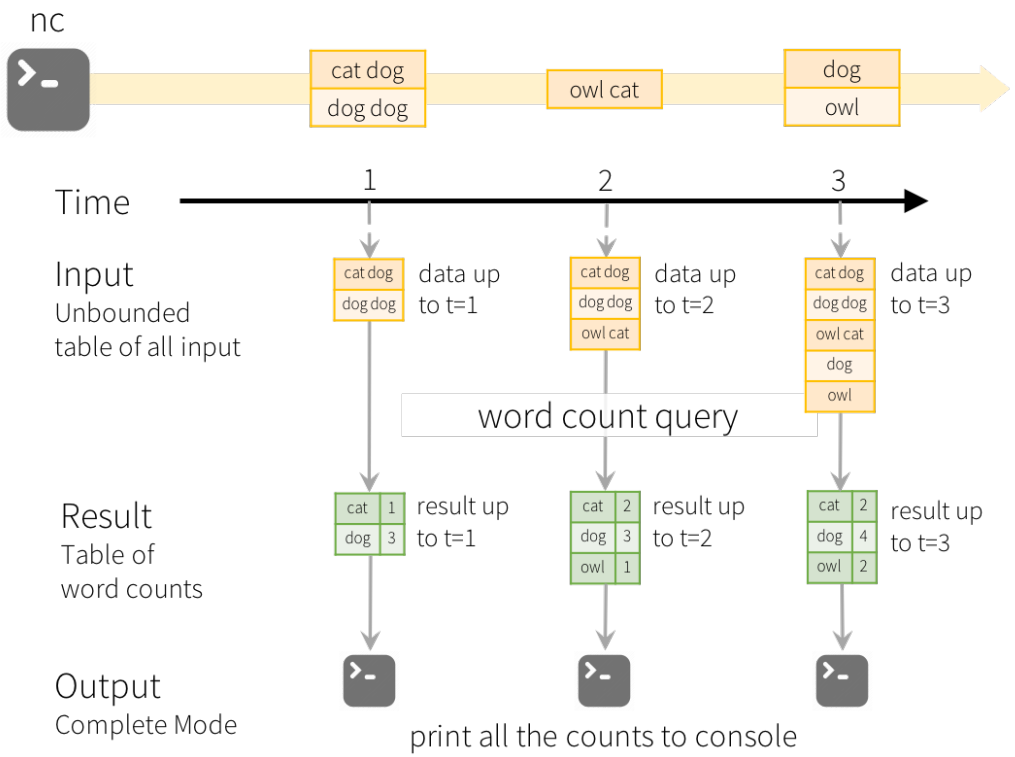
Les API de streaming sur Spark



Différence Structured Streaming et Spark Streaming

	Spark Streaming	Structured Streaming
Streaming réel	Micro-batch	Très proche
Structure de données	RDDs	Datasets - DataFrames
Données en retard	Non	Oui

Spark Structured Streaming



Model of the Quick Example

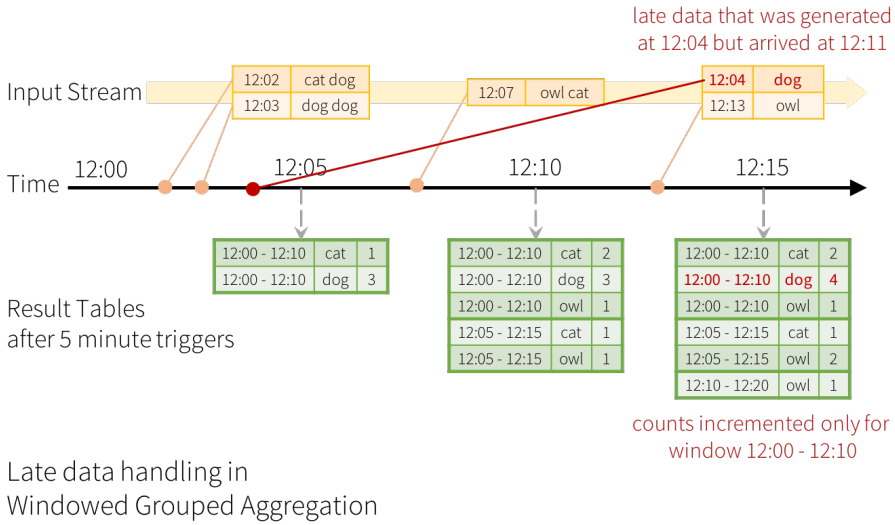
Trois modes:

- Complete
- Append
- Update

Structured Streaming ne garde pas l'agrégat des données en entrée mais va seulement garder le résultat

Spécificités de Spark Structured streaming

- Simple d'utilisation
- Peut traiter les données « en retard »
- Tolérance aux fautes



Entrées Spark Structured Streaming

- **Fichiers** : lit les fichiers arrivant dans un dossier(local, HDFS …)
- **Apache Kafka**
- **Socket (pour les tests)**
- **Rate (pour les tests)** : générateur de message à un rythme donné

Exemple: Compter les hashtags sur Twitter

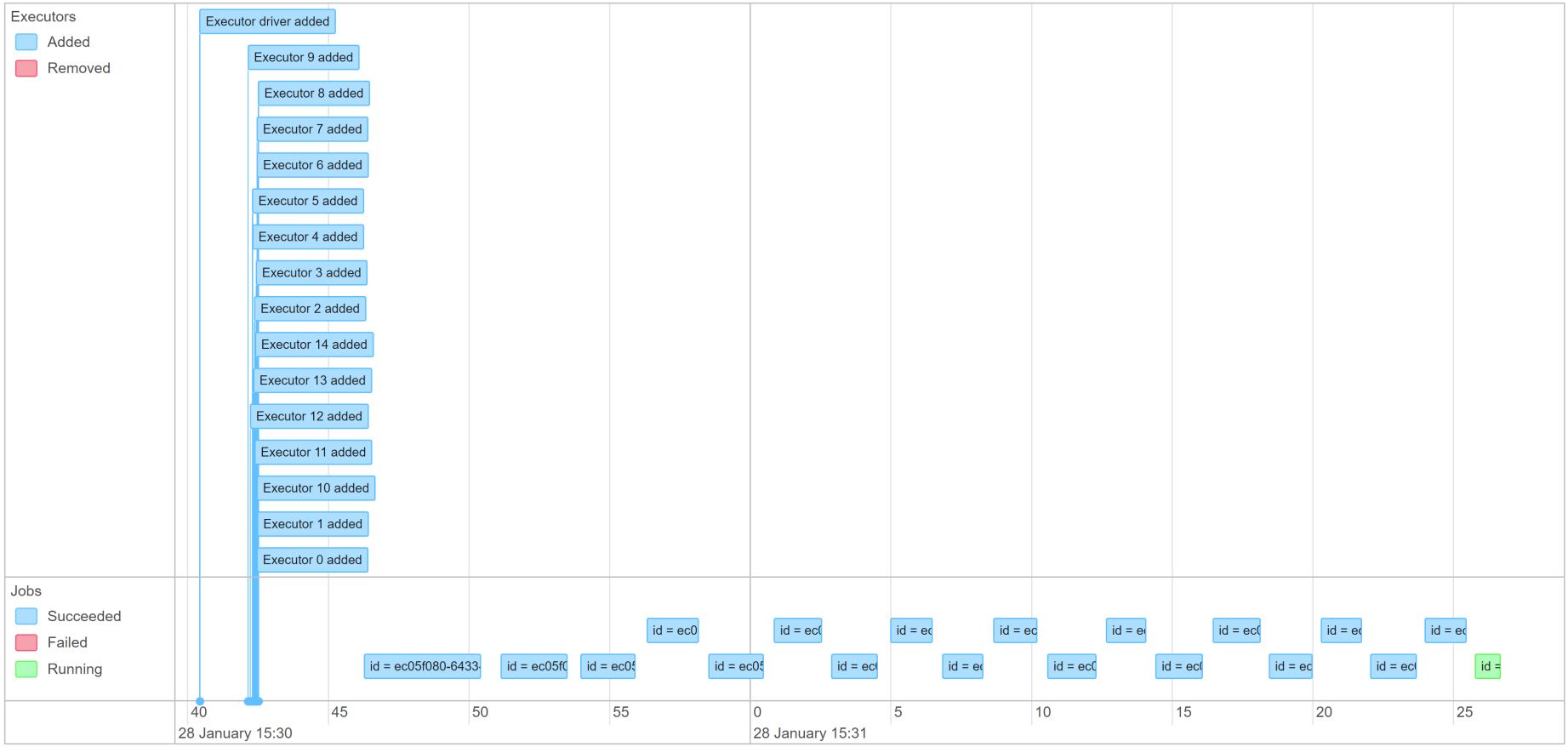
Spark Streaming

```
def process_rdd(time, rdd):
    print("----------- %s -----------" % str(time))
    try:
        sql_context = get_sql_context_instance(rdd.context)
        row_rdd = rdd.map(lambda w: Row(hashtag=w[0],
hashtag_count=w[1]))
        hashtags_df = sql_context.createDataFrame(row_rdd)
        hashtags_df.registerTempTable("hashtags")
        hashtag_counts_df = sql_context.sql(
            "select hashtag, hashtag_count from hashtags order
by hashtag_count desc limit 10")
        hashtag_counts_df.show()
    except:
        e = sys.exc_info()[0]
        print("Error: %s" % e)



words = dataStream.flatMap(lambda line: line.split(" "))
hashtags = words.filter(lambda w: '#' in w).map(lambda x: (x,
1))
tags_totals = hashtags.updateStateByKey(aggregate_tags_count)
tags_totals.foreachRDD(process_rdd)
ssc.start()
ssc.awaitTermination()
```

Structured Streaming

```
spark = SparkSession \
    .builder \
    .appName("TopHashatgs") \
    .getOrCreate()


spark.sparkContext.setLogLevel("ERROR")


lines = spark \
    .readStream \
    .format("socket") \
    .option("host", "sparkdesk.ic.metz.supelec.fr") \
    .option("port", 9009) \
    .load()


hastags = lines.select(
    explode(
        split(lines.value, " ")
    ).alias("hashtag")
).where(col("hashtag").startswith("#"))


hashtagCounts = hastags.groupBy("hashtag").count()


query = hashtagCounts \
    .writeStream \
    .outputMode("complete") \
    .format("console") \
    .start()
```

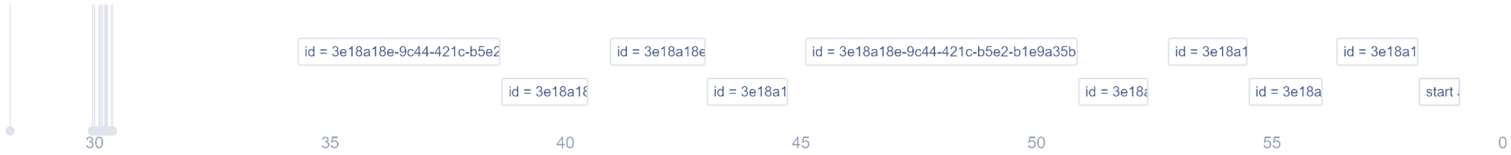
Spark UI


Executors
Added
Removed
Executor driver added
Executor 9 added
Executor 8 added
Executor 7 added
Executor 6 added
Executor 5 added
Executor 4 added
Executor 3 added
Executor 2 added
Executor 14 added
Executor 13 added
Executor 12 added
Executor 11 added
Executor 10 added
Executor 1 added
Executor 0 added
Jobs
Succeeded
Failed
Running
id = ec0
id = ec05f080-6433-
id = ec05f0
id = ec05
id = ec05
40
45
50
55
28 January 15:30
0
5
10
15
20
25
28 January 15:31

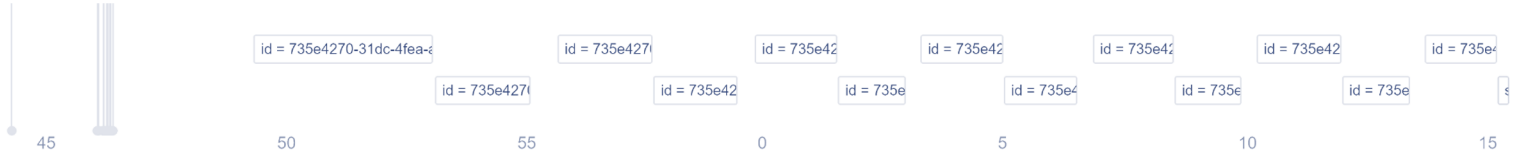
Tests de performance

On utilise une source *Rate*

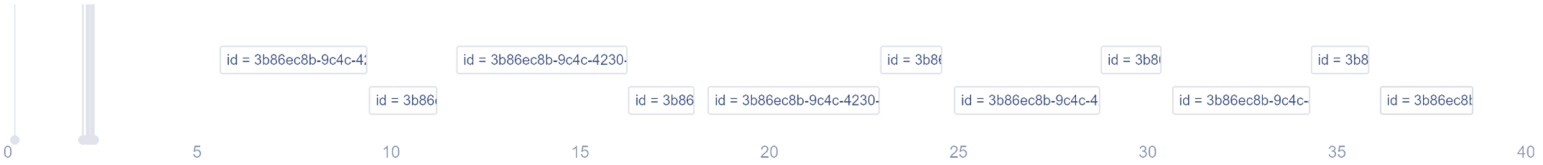
1000 /s



100 000 /s



10 000 000 /s

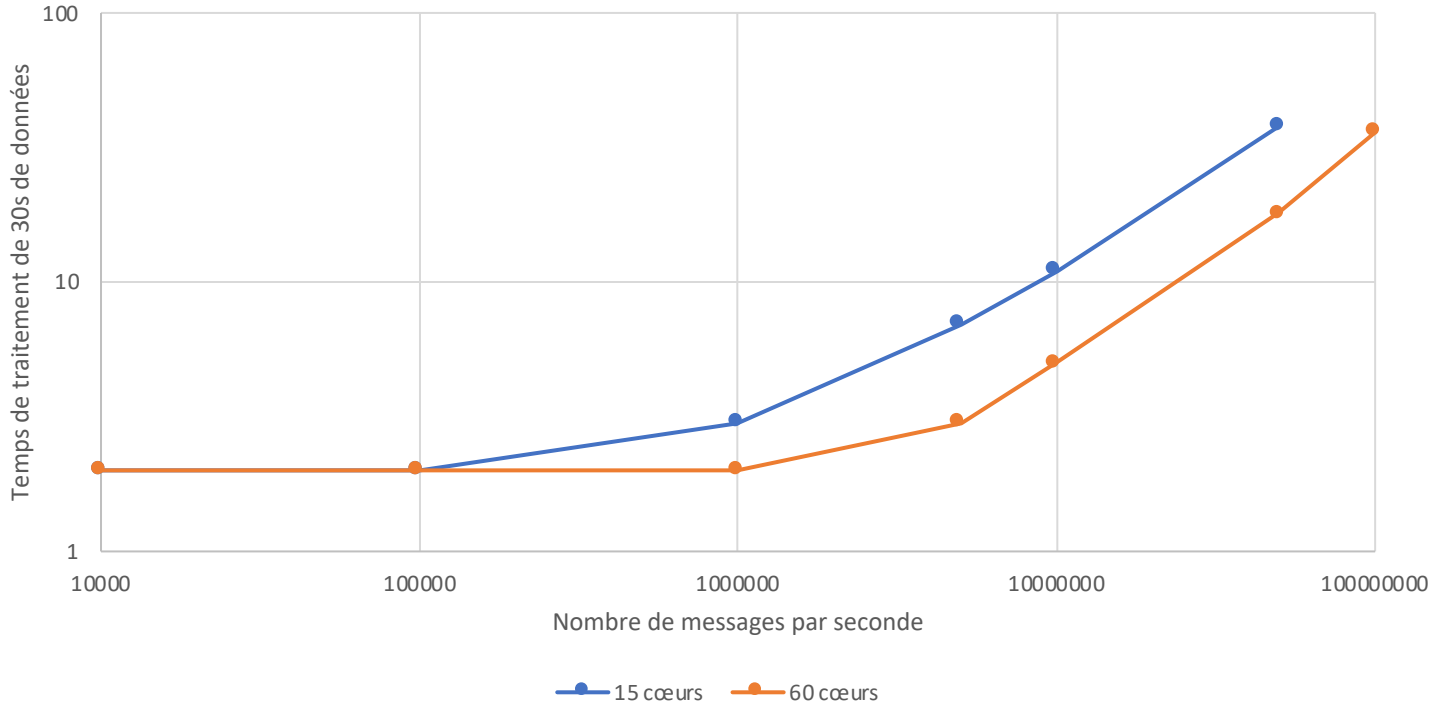


Tests de performance

100 000 000 /s
15 cœurs

100 000 000 /s
60 cœurs

Tests de performance



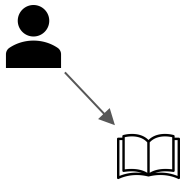
Analyse d'un graphe
Achats sur Amazon

Les données

- Simulation d’achats de produits Amazon
- Achats enregistrés sur le HDFS en csv

marketplace	customer_id	review_id	product_id	product_parent	product_title
FR	14952	R32VYUWDIB5LKE	552774294	362925721	The God Delusion
FR	14952	R3CCMP4EV6HAVL	B004GJXQ20	268067011	A Game of Thrones (A Song of Ice and Fire, Book 1)
FR	17564	R14NAE6UGTVTA2	B00GIGGS6A	256731097	Huion H610 PRO
FR	18940	R2E7QEWSC6EWFA	B00CW7KK9K	977480037	Withings Pulse - Suivi d'activitV© + Analyse du sommeil + FrV©quence cardiaque, Noir
FR	20315	R26E6I47GQRYKR	B002L6SKIK	827187473	Prometheus
FR	20842	R1RJMTSNCKB9LP	B00004YLIU	678427290	Kid A
FR	20913	R2P2XF84YELQBZ	B00AYHK7RU	108403123	The Next Day
FR	21490	RJKULSX2Y5R07	B00CJ3V5UK	252503117	G.I. Joe 2 : Conspiration [Combo Blu-ray + DVD]
FR	24196	R3UYE0U7SQCI8Q	B000FUM0TE	48021829	Amistad
FR	24196	R1TKJ7XFS3RDEB	B000UMXCTY	181554537	The Collection (Coffret 5 CD + 1 DVD)
FR	24196	R3S9JNS21QDBXP	B0044JV1K6	386772628	Braddock - Missing in action 3 [Import anglais]
FR	24196	R2Y4O06QMOGHD0	2226249672	753674225	Au revoir lV†-haut - Prix Goncourt 2013
FR	24196	R3PS3P7G1ZT57W	B000A3IF8G	757843869	The Essential Michael Jackson

Structure



Amazon Customer Review Dataset
Un client vient acheter un produit (deux nœuds, une arrête)

Streaming

Dataframe Arrêtes

Dataframe Nœuds

Opération sur le graphe avec GraphFrame:

- Connected componnents
- Label Propagation
- Page Rank

Performances