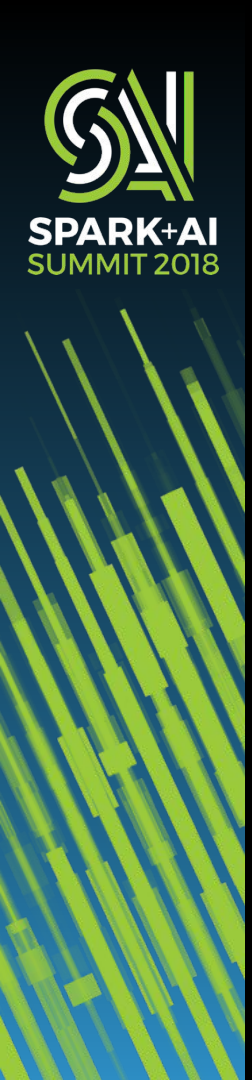


Real-Time Attribution with Structured Streaming and Databricks Delta

Caryl Yuhas, Databricks

#ExpSAIS13

Introduction

- **Goal:**
  Provide tools and information that can help you build more real-time / lower latency attribution pipelines
- Crawl, Walk, Run: Pull Model

Getting Started

- What is Attribution?

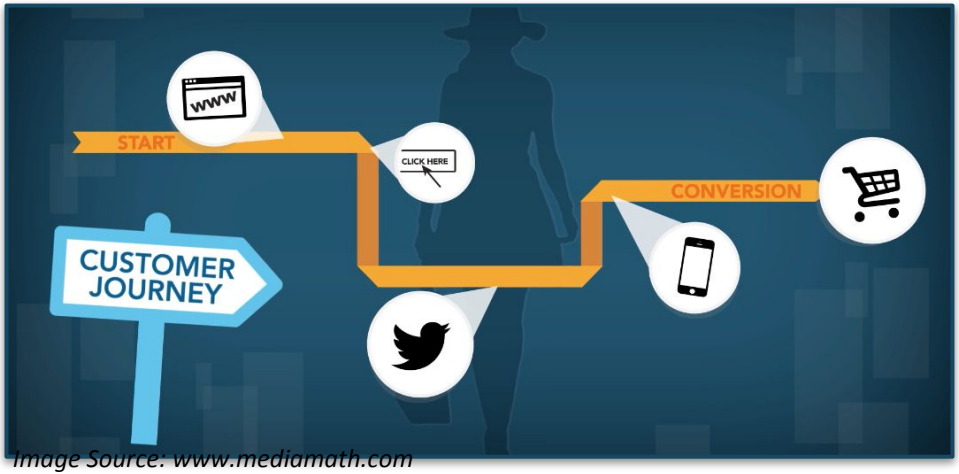


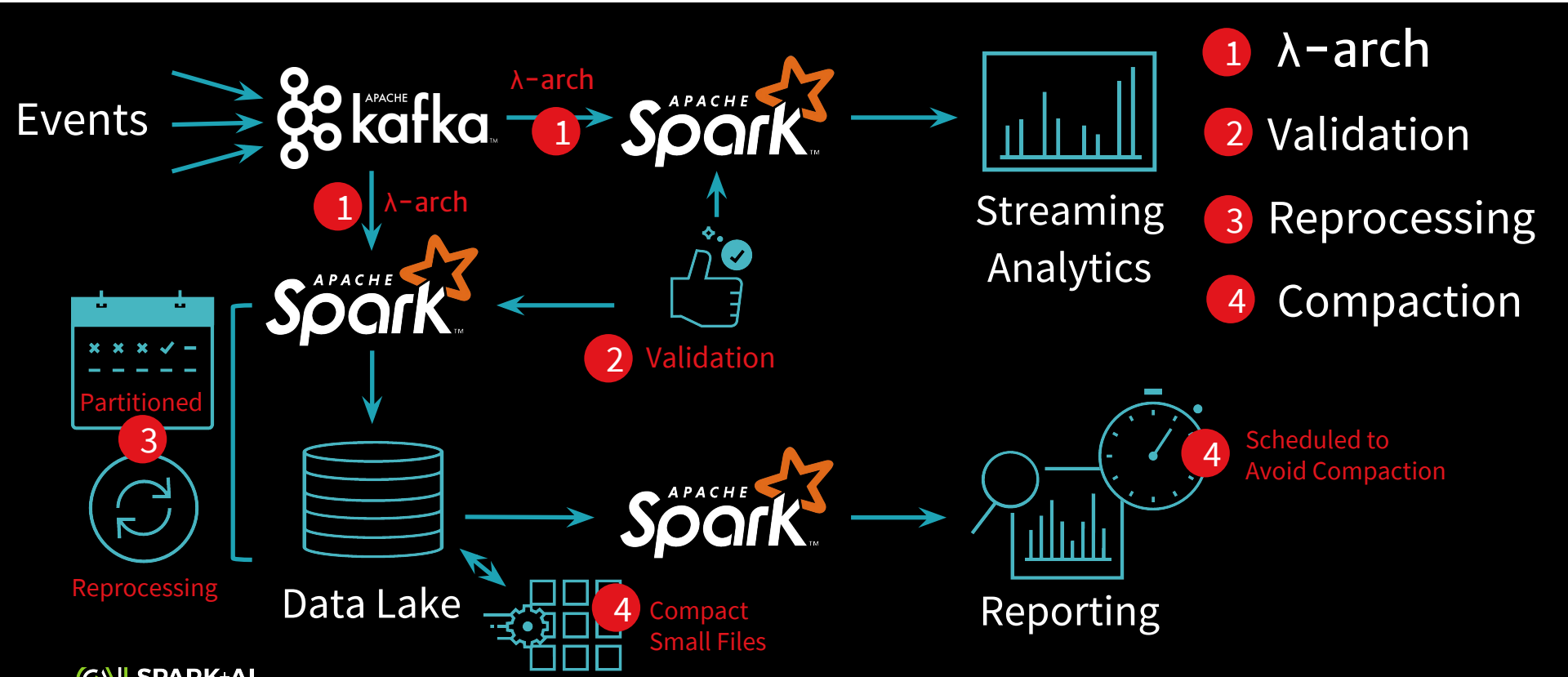
Image Source: www.mediamath.com

Introduction

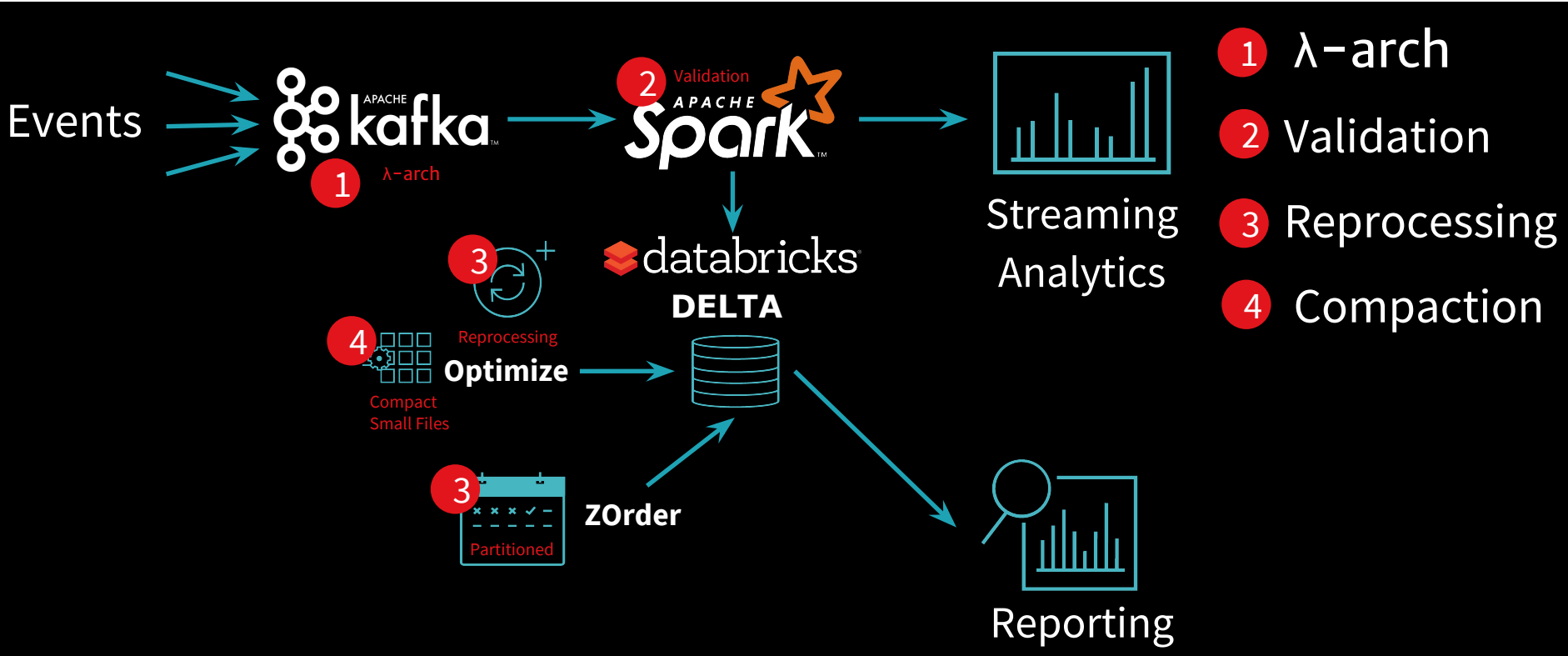
What is Databricks Delta?

Delta is a data management capability that brings data reliability and performance optimizations to the cloud data lake.

Stream-to-Sink BEFORE



Stream-to-Sink AFTER



Attribution in Practice

timestamp	exchangeID	publisher	creativeID	click	advertiserID	uid	browser	geo	bidAmount
2017-11-05T03:30:41	11	facebook.com	646594	0	523981	ca9766-h185-639bb-39d139322	Internet Explorer	Washington DC	0.33999999999999997
2017-11-01T22:19:16	9	facebook.com	248917	0	523981	ae1352-b550-108dd-08c764209	Firefox	Washington DC	0.62
2017-11-03T04:07:47	43	yahoo.com	451648	0	523981	hh6828-h529-965ge-09a995880	Chrome	Dallas	0.27
2017-11-03T19:35:14	33	facebook.com	706371	0	523981	af3869-g782-898ag-98c836867	Chrome	Houston	0.67
2017-11-03T18:27:55	36	yahoo.com	789446	1	523981	fa8434-h384-405ah-06c338074	Firefox	San Francisco	0.61

impressions

timestamp	conversionID	advertiserID	pixelID	uid	conversionValue
2018-05-24T20:19:42	618680712	702394	297606	hh8802-d498-306gc-57f819532	50.94
2018-05-24T20:19:42	733664962	489251	510749	ag9730-h993-520bb-22a473541	3.14
2018-05-24T20:19:49	717886952	489251	510749	bf2208-f421-620dc-29d253978	20.03
2018-05-24T20:19:54	681228883	489251	510749	fc5809-h538-323ha-90f379923	10.39
2018-05-24T20:19:59	815829248	523981	476019	fd6211-e549-700hc-82d403887	22.69
2018-05-24T20:20:03	618218897	702394	297606	ec4879-f243-903ab-32e202536	46.71
2018-05-24T20:20:05	445116196	523981	476019	ab3809-h114-656cd-24a029945	28.02
2018-05-24T20:20:10	709574205	702394	297606	ag4438-f396-319bc-56e125399	30.96

conversions

JOIN

attributed impressions

impTimestamp	exchangeID	publisher	creativeID	click	uid	browser	geo	bidAmount	date	convTimestamp	conversionID	advertiserID	pixelID	conversionValue	attrRank	numImps	weightedRevAttr
2017-11-05T14:54:12	31	vice.com	665594	0	ah5028-c776-845ed-93c644446	Safari	New York	0.18	2017-11-05	2018-06-03T12:11:38	123674064	523981	476019	40.02	1	1	40.02
2017-11-07T01:42:09	10	hearst.com	331008	1	fh5141-f608-233gd-70f367302	Safari	Chicago	0.7	2017-11-07	2018-05-20T16:18:44	123800875	702394	297606	17.490000000000002	1	1	17.490000000000
2018-05-12T04:51:21	17	hearst.com	498516	0	dh9446-c030-468fg-	Firefox	New York	0.54	2018-05-12	2018-05-12T11:05:02	123829322	489251	510749	42.540000000000006	1	4	10.635000000000

Attribution Challenges

Scale

- Often dealing with millions to billions of data points per attribution window

Complexity

- Simple, last-click model is still common
- MTA and more sophisticated attribution on rise

High Level Attribution Pipeline

Attribution in Practice

timestamp	exchangeID	publisher	creativeID	click	advertiserID	uid	browser	geo	bidAmount
2017-11-05T03:30:41	11	facebook.com	646594	0	523981	ca9766-h185-639bb-39d139322	Internet Explorer	Washington DC	0.33999999999999997
2017-11-01T22:19:16	9	facebook.com	248917	0	523981	ae1352-b550-108dd-08c764209	Firefox	Washington DC	0.62
2017-11-03T04:07:47	43	yahoo.com	451648	0	523981	hh6828-h529-965ge-09a995880	Chrome	Dallas	0.27
2017-11-03T19:35:14	33	facebook.com	706371	0	523981	af3869-g782-898ag-98c836867	Chrome	Houston	0.67
2017-11-03T18:27:55	36	yahoo.com	789446	1	523981	fa8434-h384-405ah-06c338074	Firefox	San Francisco	0.61

timestamp	conversionID	advertiserID	pixelID	uid	conversionValue
2018-05-24T20:19:42	618680712	702394	297606	hh8802-d498-306gc-57f819532	50.94
2018-05-24T20:19:42	733664962	489251	510749	ag9730-h993-520bb-22a473541	3.14
2018-05-24T20:19:49	717886952	489251	510749	bf2208-f421-620dc-29d253978	20.03
2018-05-24T20:19:54	681228883	489251	510749	fc5809-h538-323ha-90f379923	10.39
2018-05-24T20:19:59	815829248	523981	476019	fd6211-e549-700hc-82d403887	22.69
2018-05-24T20:20:03	618218897	702394	297606	ec4879-f243-903ab-32e202536	46.71
2018-05-24T20:20:05	445116196	523981	476019	ab3809-h114-656cd-24a029945	28.02
2018-05-24T20:20:10	709574205	702394	297606	ag4438-f396-319bc-56e125399	30.96

impressions

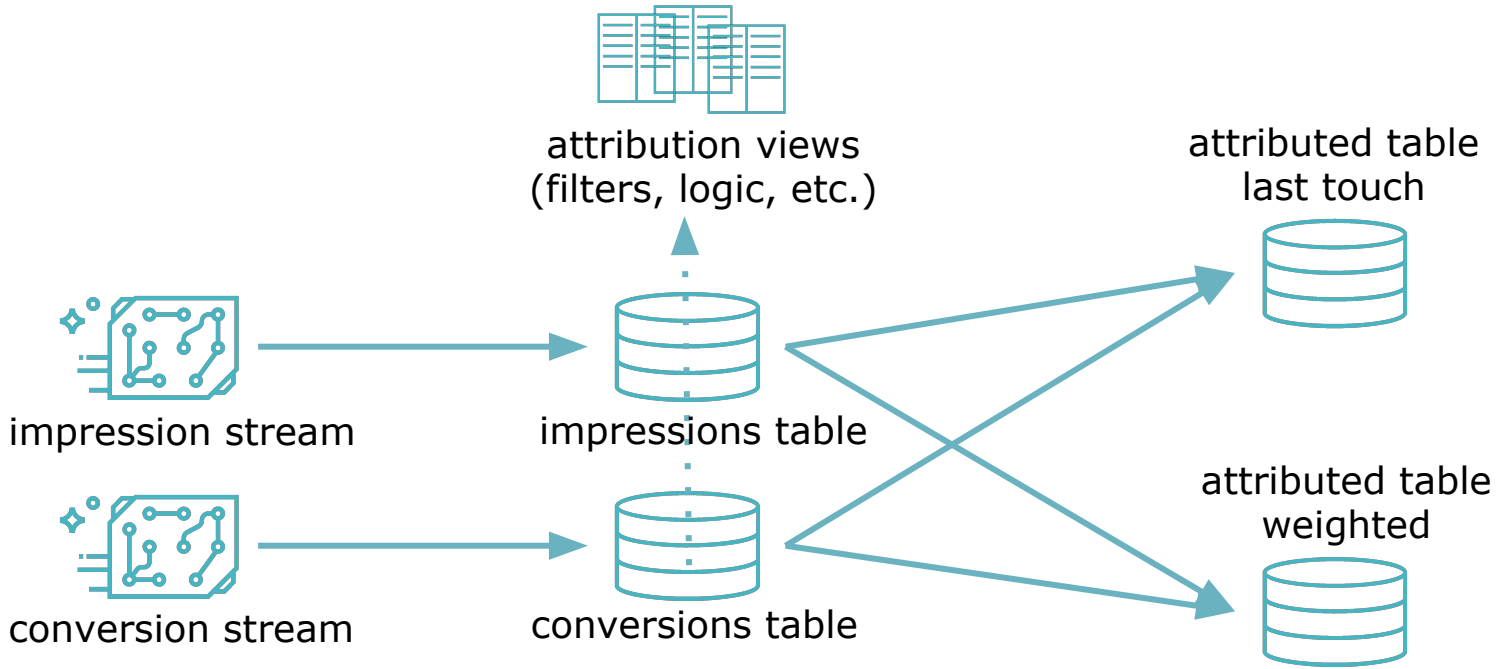
JOIN

conversions

attributed impressions

impTimestamp	exchangeID	publisher	creativeID	click	uid	browser	geo	bidAmount	date	convTimestamp	conversionID	advertiserID	pixelID	conversionValue	attrRank	numImps	weightedRevAttr
2017-11-05T14:54:12	31	vice.com	665594	0	ah5028-c776-845ed-93c644446	Safari	New York	0.18	2017-11-05	2018-06-03T12:11:38	123674064	523981	476019	40.02	1	1	40.02
2017-11-07T01:42:09	10	hearst.com	331008	1	fh5141-f608-233gd-70f367302	Safari	Chicago	0.7	2017-11-07	2018-05-20T16:18:44	123800875	702394	297606	17.490000000000002	1	1	17.490000000000
2018-05-12T04:51:21	17	hearst.com	498516	0	dh9446-c030-468fg-	Firefox	New York	0.54	2018-05-12	2018-05-12T11:05:02	123829322	489251	510749	42.540000000000006	1	4	10.635000000000

Data Architecture



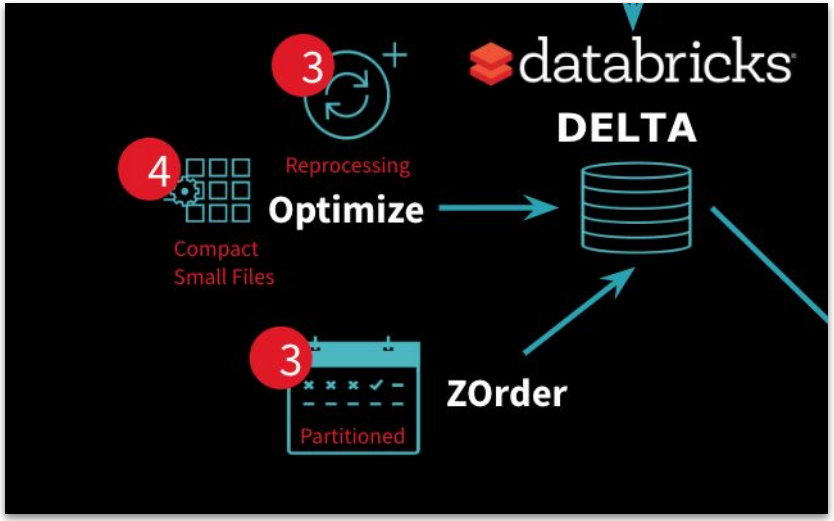
System Architecture

Unification of Streaming + Batch

DEMO

Managing Performance

- How can we optimize performance?
- **Levers:**
  - Delta Tools
    - Optimize
    - ZOrder
    - Caching
    - Data Skipping
  - Join on Stream
  - Cluster Size



Handling Complexity

- Flexibility with Complex Logic
  - Forking streams
  - Logic on query vs. in-stream
- Late or Corrected Data
  - Upserts
  - Views automatically update when raw data changed

Conclusion

- Unification of Batch & Streaming
- Easy APIs for Managing Performance
- Flexible and Scalable Analytics on Near Real-Time Data