Crafting Unbreakable Data Pipelines with Snowflake

Darren Gardner
Big Data Practice Lead

Decisive Data

Decisive Data
insights for decisions

We help people use data to realize better outcomes.

Data Apps	Data Platform	Data Analytics	Data Science

snowflake

Solution Partner

Struggle

Broken Data Pipelines

Causes

Infrastructure	• Power, network • Disk, memory
Schema drift	• Upstream database changes • Application upgrades
File formats	• Lack of consistency • Unexpected data
API	• Versioning • Switch to new API

Potential Impact

Stale Data

Lost Data

Failed State

A Shift in Thinking…

Favor guaranteed acquisition over structured data

Loss of convenience of structured data is mitigated by native semi-structured support in Snowflake SQL

ETL: Traditional	▦	Structured	→ Structured	▦
ETL: Big Data	{ : }	Semi-structured	→ Structured	▦
ELT: Modern	{ : }	Semi-structured	→ Semi-structured	{ : }
tELT?	▦	Structured	→ Semi-structured	{ : }

Solution

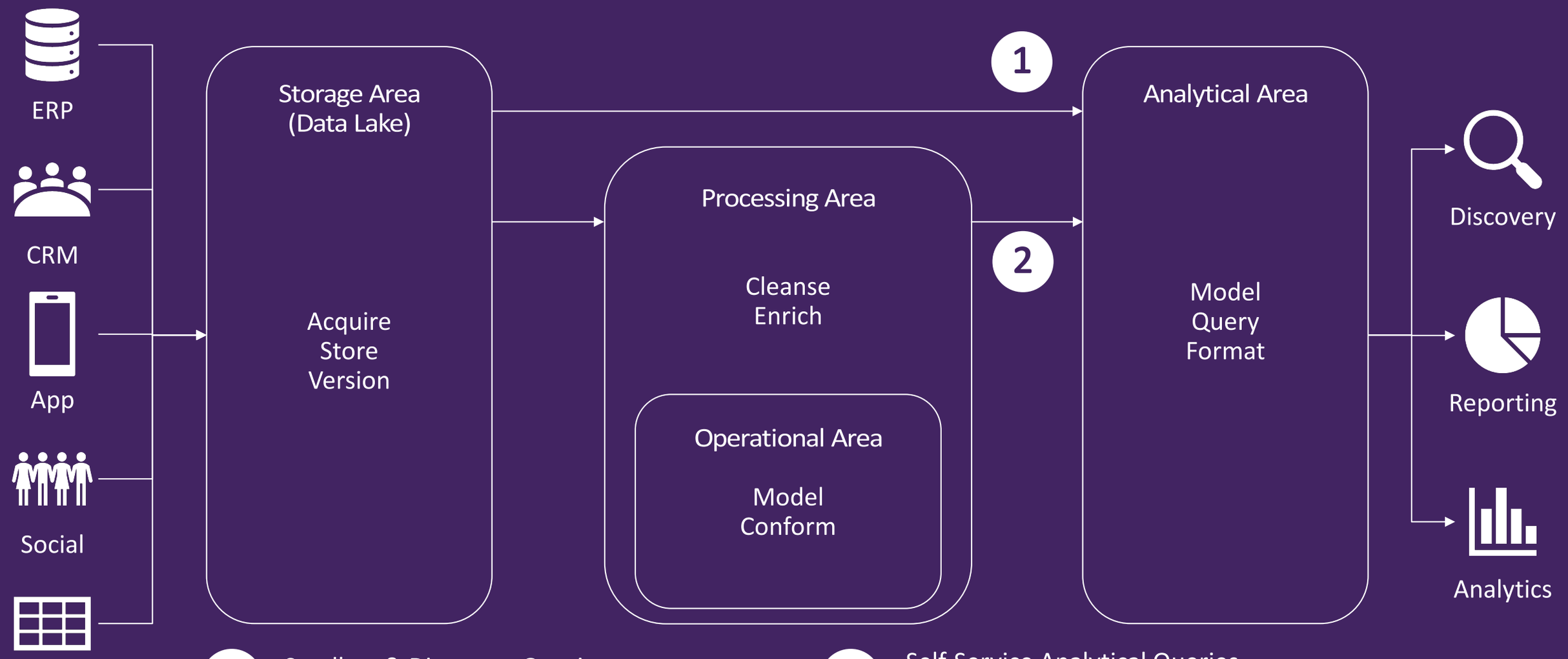
Defer analytical constraints from data acquisition and into modeling

Decouple raw data acquisition from ingestion into analytic models

Immunize the pipeline from the impact of schema drift

Convert structured formats to semi-structured JSON

Architecture



1 Sandbox & Discovery Queries

Optimized for Fast & Quick Data Discovery

2 Self-Service Analytical Queries

Optimized for High Quality, Easy-to-use Self Service Analysis

Example: Data

Structured Format

F_Name	L_Name	Pos	Age
Robinson	Cano	2B	34
Kyle	Seager	3B	29
Nelson	Cruz	DH	36

JSON Equivalent


```
{"Age":"34","F_Name":"Robinson","L_Name":"Cano","Pos":"2B"}
{"Age":"29","F_Name":"Kyle","L_Name":"Seager","Pos":"3B"}
{"Age":"36","F_Name":"Nelson","L_Name":"Cruz","Pos":"DH"}
```

Generic Raw Data Capture Table Format

```
--Full Document
{
   "Age":"34"
  ,"F_Name":"Robinson"
  ,"L_Name":"Cano"
  ,"Pos":"2B"
}
```

SHA1()

```
--Key Document
{
  "F_Name":"Robinson"
  ,"L_Name":"Cano"
}
```

SHA1()

JSON_DOC	JSON_DOC_HASH	JSON_KEY	JSON_KEY_HASH	ATTRS
{ "Age": "34", "F_Name": "Robinson", "L_Name": "Cano", "Pos": "2B" }	7bab66c49443f7…	{ "F_Name": "Robinson", "L_Name": "Cano", }	f59fa347ee0eb05c4…	{ "LOAD_DT": "2017-03-04", "SRC_FILE": "SEA.TXT" }

Example: Schema Drift

Age → Birthdate

{**"Birthdate":"1982-10-22",**"F_Name":"Robinson","L_Name":"Cano","Pos":"2B"}
{**"Birthdate":"1987-11-03",**"F_Name":"Kyle","L_Name":"Seager","Pos":"3B"}
{**"Birthdate":"1980-07-01",**"F_Name":"Nelson","L_Name":"Cruz","Pos":"DH"}

JSON_DOC	JSON_DOC_HASH	JSON_KEY	JSON_KEY_HASH	ATTRS
{ "Age": "34", "F_Name": "Robinson", "L_Name": "Cano", "Birthdate": "1982-10-22" }	0173A349A53FAC2…	{ "F_Name": "Robinson", "L_Name": "Cano", }	f59fa347ee0eb05c4…	{ "LOAD_DT": "2017-04-17", "SRC_FILE": "SEA.TXT" }

Historical, Versioned Data

JSON_DOC	JSON_DOC_HASH	JSON_KEY	JSON_KEY_HASH	START_DT	END_DT	ATTRS
{ "Age": "34", "F_Name": "Robinson", "L_Name": "Cano", "Pos": "2B" }	7bab66c49443f7…	{ "F_Name": "Robinson", "L_Name": "Cano", }	f59fa347ee0eb05c4…	2017-03-04 00:00:00	2017-04-17 00:00:00	{ "LOAD_DT": "2017-03-04", "SRC_FILE": "SEA.TXT" }
{ "Age": "34", "F_Name": "Robinson", "L_Name": "Cano", "Birthdate": "1982-10-22" }	0173A349A53FAC2…	{ "F_Name": "Robinson", "L_Name": "Cano", }	f59fa347ee0eb05c4…	2017-04-17 00:00:00	9999-12-31 00:00:00	{ "LOAD_DT": "2017-04-17", "SRC_FILE": "SEA.TXT" }

Example: Analytical Query Modeling

```
SELECT JSON_DOC:"F_Name"          ::STRING        AS F_Name
  ,JSON_DOC:"L_Name"              ::STRING        AS L_Name
  ,JSON_DOC:"Pos"                 ::STRING        AS Pos
  ,JSON_DOC:"Age"                 ::INTEGER       AS Age_Raw
  ,JSON_DOC:"Birthdate"           ::DATE  AS Birthdate
  ,COALESCE(JSON_DOC:"Age"::INTEGER, DATEDIFF(YEAR, JSON_DOC:"Birthdate"::DATE, CURRENT_DATE) - CASE
WHEN DATE_FROM_PARTS(DATE_PART(YEAR, CURRENT_DATE), DATE_PART(MONTH, JSON_DOC:"Birthdate"::DATE),
DATE_PART(DAY, JSON_DOC:"Birthdate"::DATE)) > CURRENT_DATE THEN 1 ELSE 0 END) AS Age
  ,START_DT,END_DT
FROM    PSA.PLAYER;
ORDER BY F_Name, L_Name, START_DT
```

F_NAME	L_NAME	POS	AGE_RAW	BIRTHDATE	AGE	START_DT	END_DT
Kyle	Seager	3B	29	NULL	29	2017-03-04 00...	2017-04-17 00..
Kyle	Seager	3B	NULL	1987-11-03	29	2017-04-17 00...	9999-12-31 00..
Nelson	Cruz	DH	36	NULL	36	2017-03-04 00...	2017-04-17 00..
Nelson	Cruz	DH	NULL	1980-07-01	36	2017-04-17 00...	9999-12-31 00..
Robinson	Cano	2B	34	NULL	34	2017-03-04 00...	2017-04-17 00..
Robinson	Cano	2B	NULL	1982-10-22	34	2017-04-17 00...	9999-12-31 00..

Darren Gardner
Big Data Practice Lead
darren@decisivedata.net