ETL Best Practices & Techniques

Datalere
Empowering the Digital Enterprise

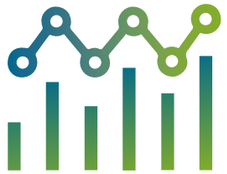
ABOUT THE PRESENTER

Marc Beacom

Managing Partner
Business Intelligence

- More than 15 years of experience in Microsoft data engineering and business intelligence
- mbeacom@Datalere.com
- @MarcBeacom

ABOUT DATALERE



Data Science

Using Azure to Embrace
your Most Valuable Insights

Data Engineering

Modern Data
Management Solutions

Emerging Technologies

Advanced Architectures and
Machine Learning

Business Intelligence

Enable a Data-Driven
Organization

Data Architecture

Enable Fast and Easy Access
to your Organization's Data

Application & Mobile Development

Solutions to Support Your
Unique Business Needs

CLOUD PARTNERS

Microsoft Partner
Microsoft
Gold Data Analytics

amazon web services | Partner Network

Google Cloud Platform
Partner

OUR CUSTOMERS AND PROJECTS

POLL – ARE YOU AN…

- ETL Engineer ?
- Database Engineer ?
- DBA ?
- Manager / Director ?
- Other ?

#01: ETL TEMPLATES

- Start developing right away with a known and consistent framework
- Pre-built with auditing/logging – just copy and paste to reuse the template
- Logging
    - Batch/Package/Table, Start/End dates for durations, Load Windows
- Auditing
    - Row counts, log rows with batch ID, rollback if needed
- Template Examples
    - Master / Parent Package, Child, Loading flat files

ETL Templates

DEMO

Datalere
Empowering the Digital Enterprise

#02 - STANDARDS

- Leverage system variables where possible such as logging
- Assists in troubleshooting
- Create and Follow Naming Standards
- Checklists for
  - Code reviews
  - Environment setup – servers and development stations

#02 -STANDARDS: CHECKLISTS

Review Checklist

The following check list is used when reviewing an ETL package prior to promoting it from Dev to QA.

Control Flow

1. The major and/or minor versions have been adjusted accordingly
2. Precedence Constraint exists after the “SQL Load Check Status” task
3. All Control Flow tasks are enabled
4. All Control Flow tasks have the proper prefix
5. Project configurations being used

Data Flow

6. All Data Flow transformations have the proper prefix
7. Data Flow matches a documented Data Flow Pattern
    a. If the Data Flow does not match a Data Flow Pattern, an annotation should be added to explain why and what the unique reason is
8. Data flow does NOT contain blocking transformations – Unless a valid reason and donuts are donated to the team prior to code review
9. Lookup transformations are set to partial cache unless no cache is needed.

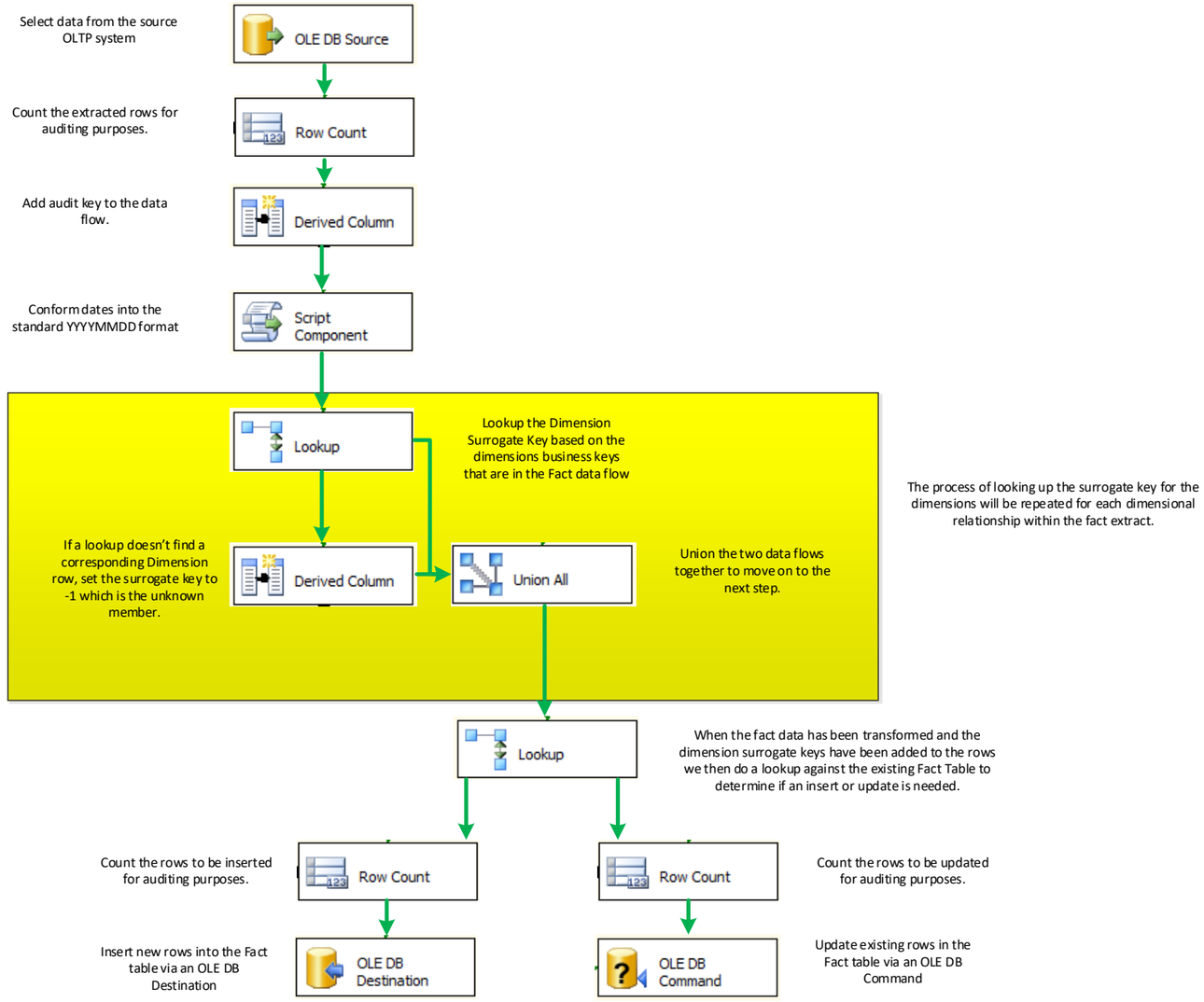
Event Handlers

10. The SQL Log Error task exists at the Package level for the OnError Event Handler.
11. All tasks in the Event Handlers are enabled and fully functional.

#03 - ETL LOAD PATTERNS

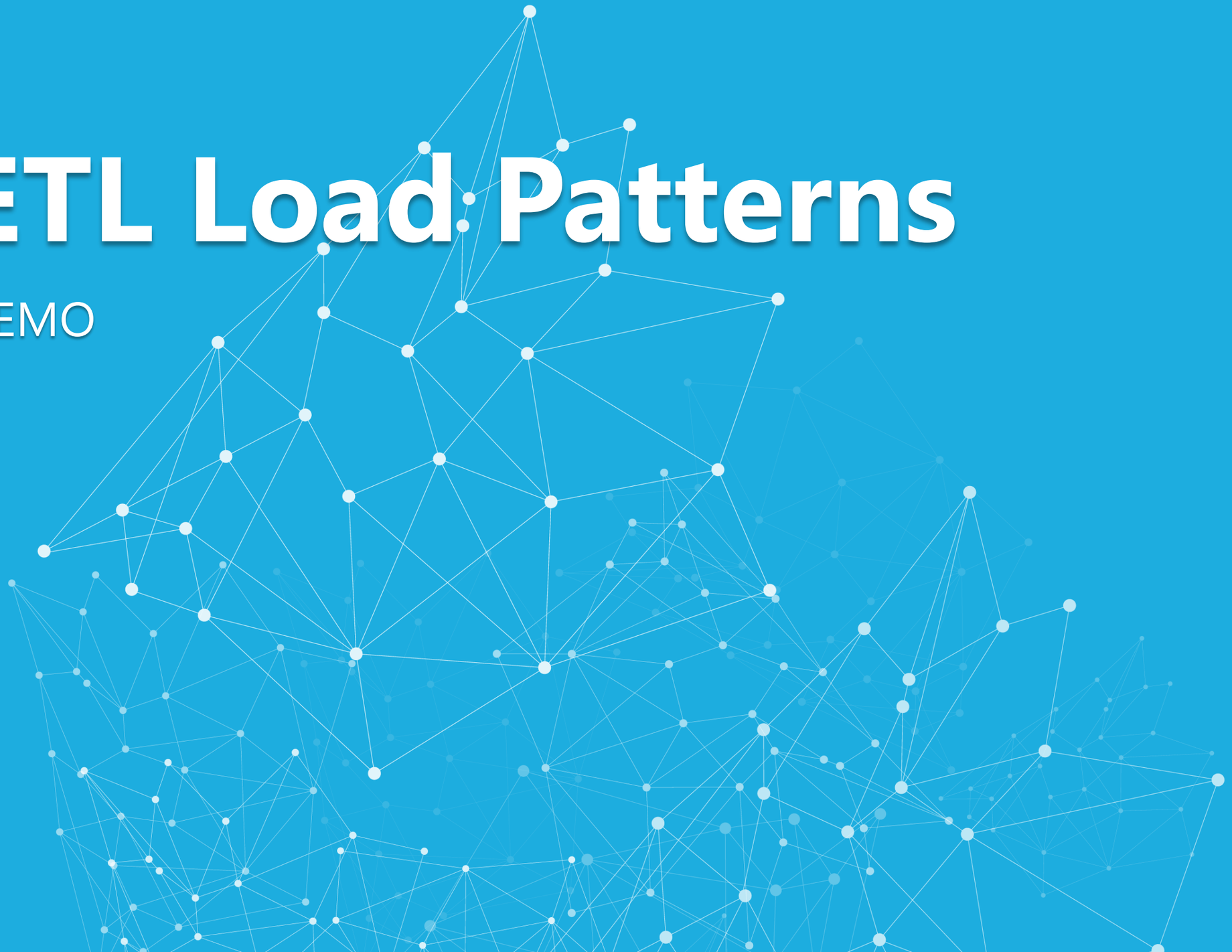
- Your ‘Blueprint’ for ETL development
- Ensures consistency when developing packages
- Reduces long-term maintenance costs
- 3-4 ETL patterns for a typical Dimensional Model
- Socialize patterns with all ETL developers

#03 - ETL LOAD PATTERN EXAMPLE



ETL Load Patterns

DEMO



Datalere
Empowering the Digital Enterprise

#04 - DOCUMENTATION

- ETL, while visual, isn’t self-documenting
- High level and not line item documentation
- Gets new team members up to speed quicker
- Include patterns, templates, standards, checklists, etc.

#04 - DOCUMENTATION

Table of Contents

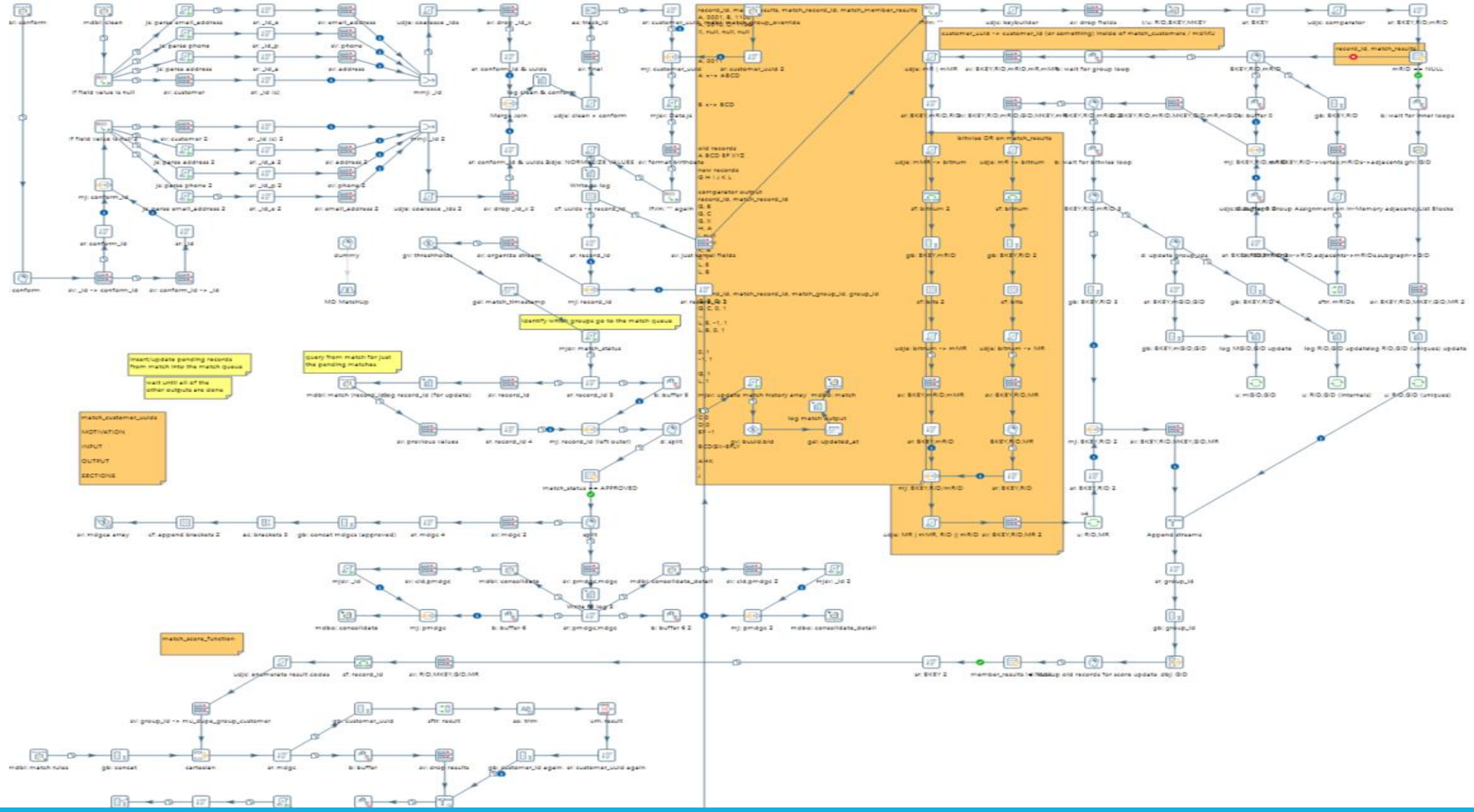
#05 - ADDRESS BAD DATA

- What is bad data? Who should define this? = You and Business!
- Develop a process, either manual or automated, to address bad data
- The outcome should be standardized and documented
- Options
  - Ignore or discard
  - Insert and flag
  - Redirect to another table/object

#06 - SIMPLIFY DATA MOVEMENT

- Move data as little as possible
- Land data close to destination
- Shortens loads cycles
- Easier to test and validate that you have met the finish line

CASE STUDY: TRANSFORMATION OVERLOAD



#06 – DECOUPLE & PERFORMANCE CASE STUDY

Challenges

- Education Data Management ISV could not successfully match and load data
- Data matching and loading often took a day to process
- Need to refactor and optimize current ETL

Recommendation

- Defined a load pattern that simplified the overall process
- Data cache adjustments allowing quicker lookups
- Sorting and Joining on database

#07 - FORMAT AND ORGANIZE

- Keep things simple but more may be less
- Add comments / annotations where needed
- Follow team standards

Formatting

DEMO

Datalere
Empowering the Digital Enterprise

#08 - INCREMENTAL LOADING

- Reduces bandwidth and times during loads – data sizes are growing!
- Achieve near real-time data refreshes – up to 2 minutes
- Consider restart ability and the ability to process a larger window
- Load only changed data and NO more

#09 - PARALLEL PROCESSING

- Take advantage of idle resources
- Control flow
  - MaxConncurrentExecutables property
  - Default of -1 = Processor count + 2
  - My Default = Processor count - 2
- Data flow
  - EngineThreads property
  - Default of 10
  - Don’t overload the server resources

#09 - PARALLEL PROCESSING CASE STUDY

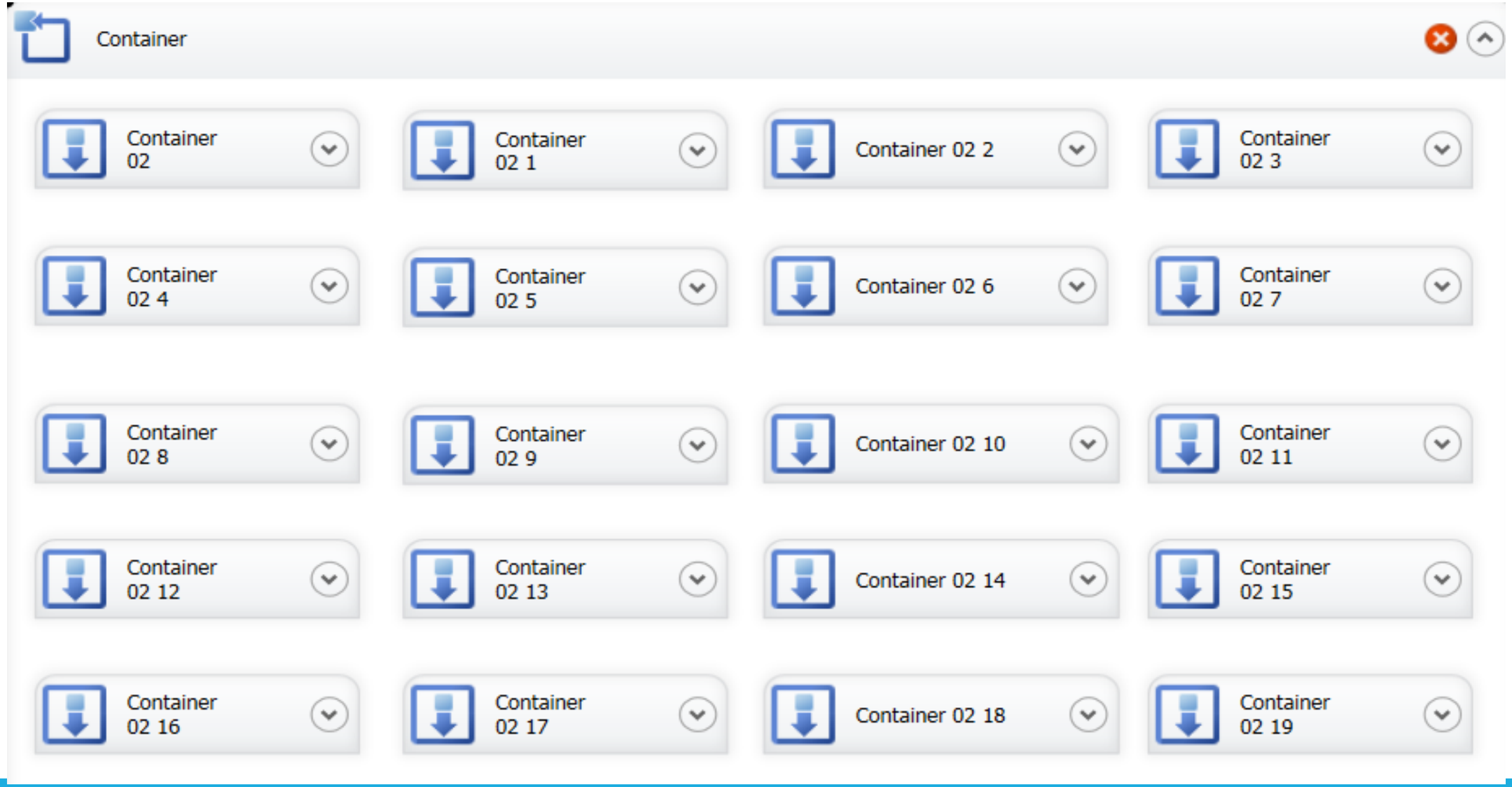
Challenges

- 1 Billion daily rows
- Current load was 6.5 hours but had 3 Hour load window
- Consumed a large percentage of resources while processing

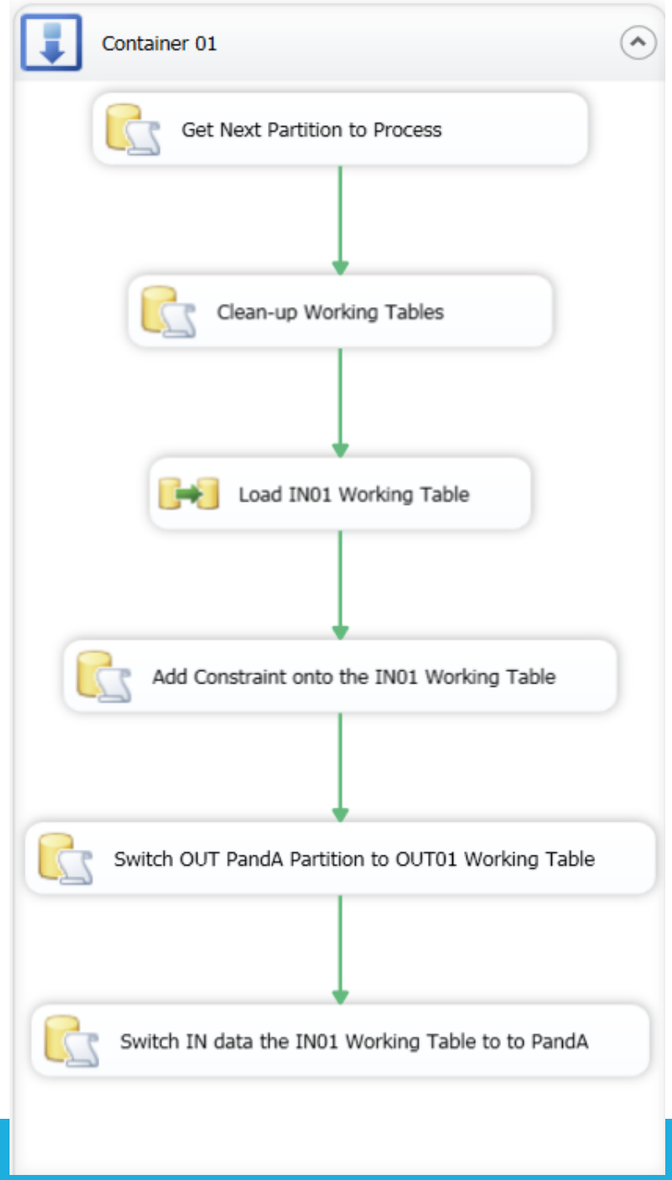
Recommendation

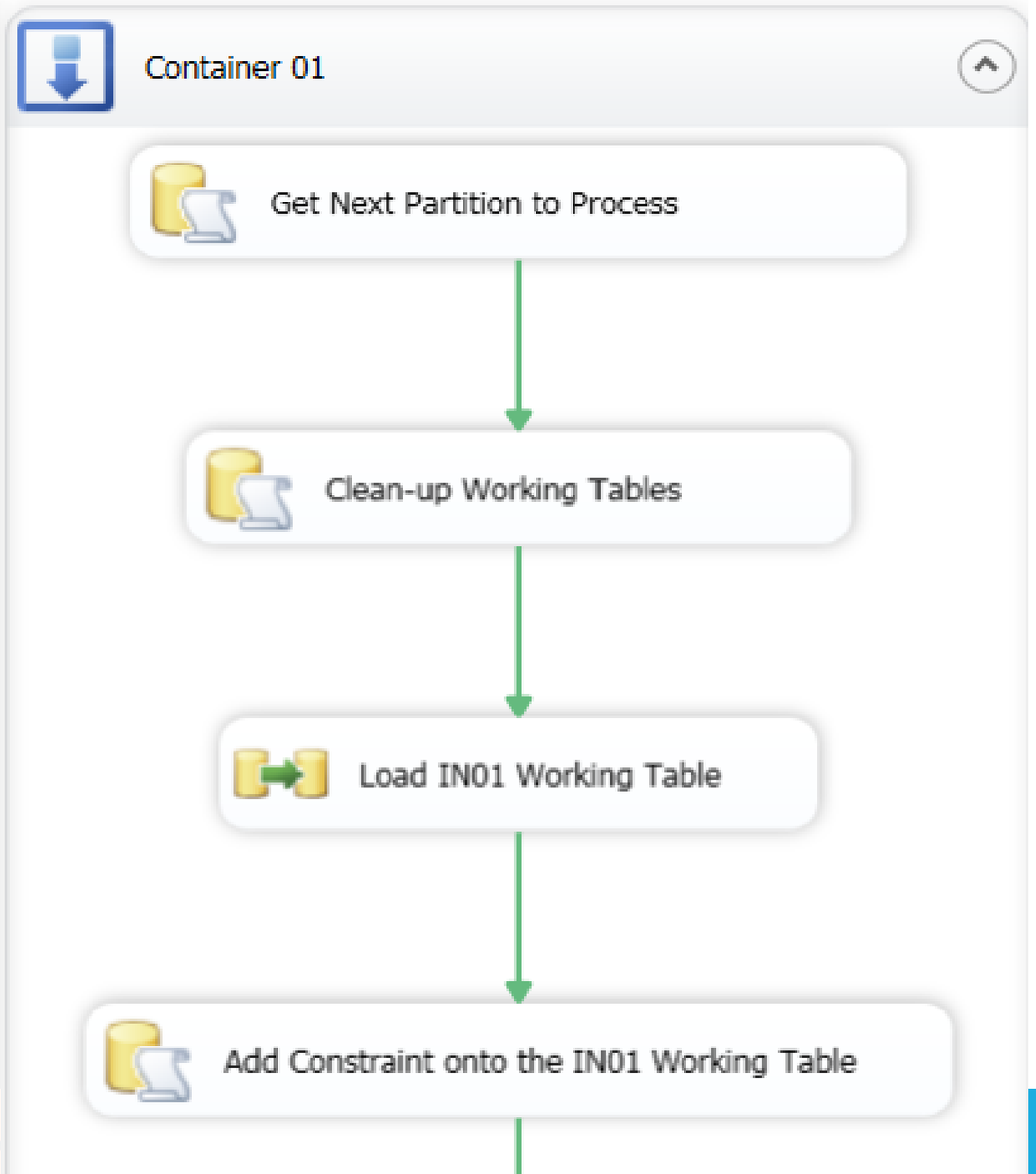
- Partitioned table
- Parallel processing – Control flow
- Better data types – GUID to BigInt – Saved 12 GB / column
- Page level compression
- Reduced load times to 2 hours

#09 - PARALLEL PROCESSING



#09 - PARALLEL PROCESSING

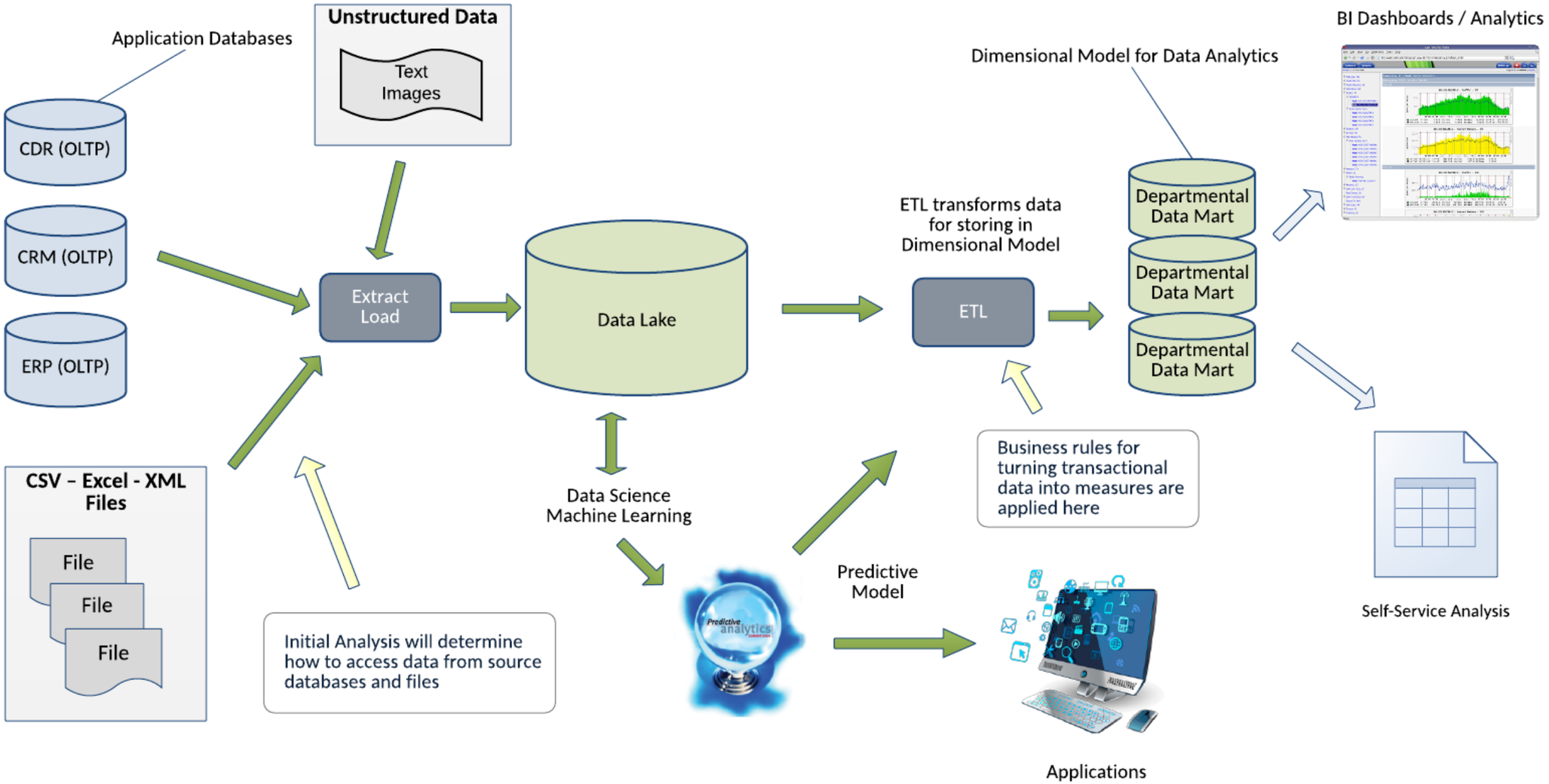




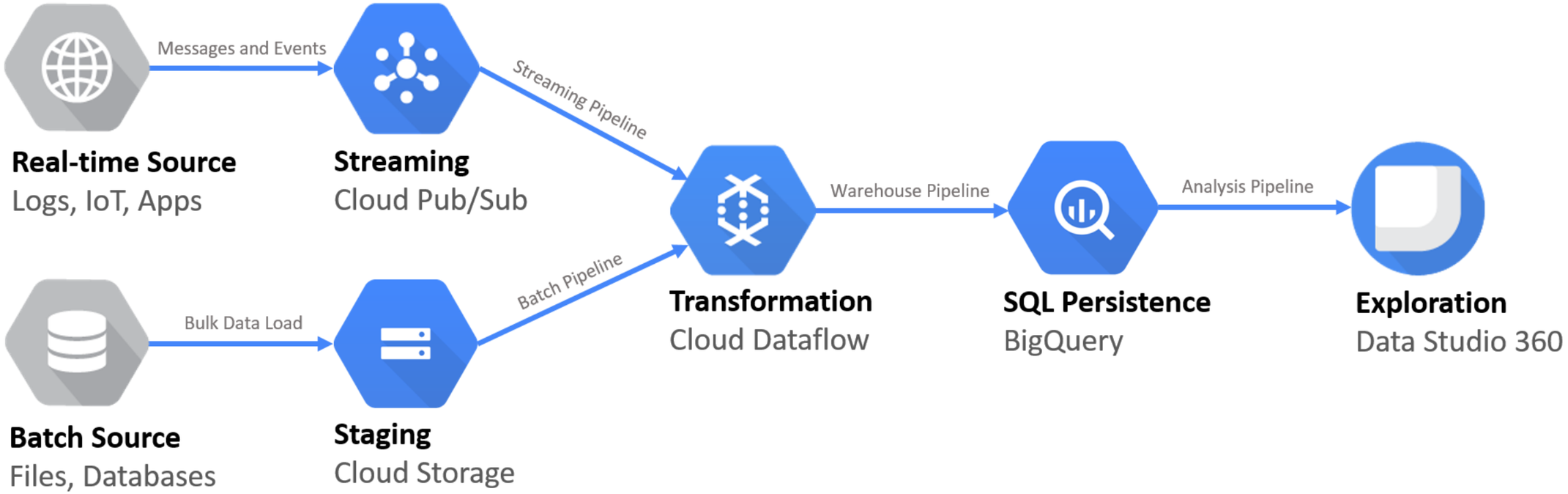
#10 - STAGE DATA

- Reduces overall ETL complexity
- Little to no transformations out of source
- Reduce the ‘hit’ on the source system
- Separate database where possible and different schema if not

DATA LAKE REFERENCE ARCHITECTURE



SIMPLIFIED SAMPLE ARCHITECTURE



#11 - OPTIMIZE SOURCE AND DESTINATIONS

- Only pull in columns you need
- Only pull in data rows you need – add a where clause
- Don’t use the select from table or view, always specify columns
- Index optimization – don’t forget about lookup queries
- Distributing data across multiple source flat files
- Sorting & Join in SQL – Sort Transformation is blocking!

#12 - BLOCKING TRANSFORMATIONS

Blocking

- Fuzzy Grouping/Lookup
- Aggregate
- Sort

Partially Blocking

- Merge Join
- Union All
- Lookup

Non-Blocking

- Derived Column
- Data Conversion
- Row Count

Blocking

DEMO

Datalere
Empowering the Digital Enterprise

#13 – VALIDATION FRAMEWORK

- Test data to ensure accuracy
- Developed my first validation framework in 2009 with SSIS
- Start small and build upon it
  - Row counts
  - Aggregations
  - Compare known data with Data Warehouse data
  - Compare source and DW data
- Can add a significate cost in development

#14 – BONUS

- Know your requirements and where the finish line is
- Build a development plan which includes testing!
- Source Control check-in often and at least daily

Thank You

Marc Beacom, Managing Partner – Business Intelligence Practice Lead
MBeacom@Datalere.com
720.319.6122

www.Datalere.com