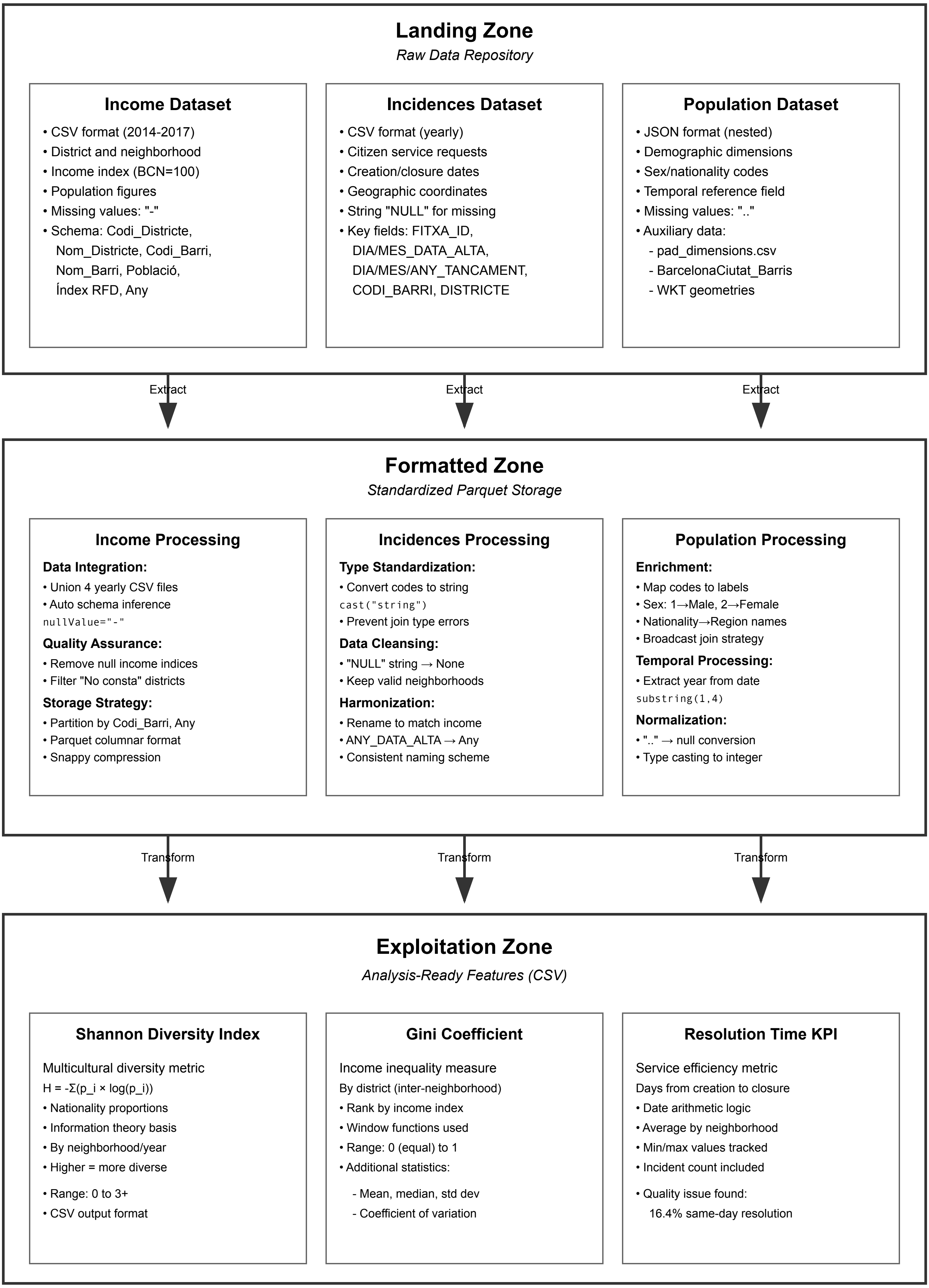

Landing Zone

Raw Data Repository

Income Dataset

- CSV format (2014-2017)
- District and neighborhood
- Income index (BCN=100)
- Population figures
- Missing values: "-"
- Schema: Codi_Districte, Nom_Districte, Codi_Barri, Nom_Barri, Població, Índex RFD, Any

Incidences Dataset

- CSV format (yearly)
- Citizen service requests
- Creation/closure dates
- Geographic coordinates
- String "NULL" for missing
- Key fields: FITXA_ID, DIA/MES_DATA_ALTA, DIA/MES/ANY_TANCAMENT, CODI_BARRI, DISTRICTE

Population Dataset

- JSON format (nested)
- Demographic dimensions
- Sex/nationality codes
- Temporal reference field
- Missing values: ".."
- Auxiliary data:
  - pad_dimensions.csv
  - BarcelonaCiutat_Barris
  - WKT geometries

Formatted Zone

Standardized Parquet Storage

Income Processing

Data Integration:

- Union 4 yearly CSV files
- Auto schema inference

`nullValue="-"`

Quality Assurance:

- Remove null income indices
- Filter "No consta" districts

Storage Strategy:

- Partition by Codi_Barri, Any
- Parquet columnar format
- Snappy compression

Incidences Processing

Type Standardization:

- Convert codes to string

`cast("string")`

- Prevent join type errors

Data Cleansing:

- "NULL" string → None
- Keep valid neighborhoods

Harmonization:

- Rename to match income
- ANY_DATA_ALTA → Any
- Consistent naming scheme

Population Processing

Enrichment:

- Map codes to labels
- Sex: 1→Male, 2→Female
- Nationality→Region names
- Broadcast join strategy

Temporal Processing:

- Extract year from date

`substring(1,4)`

Normalization:

- ".." → null conversion
- Type casting to integer

Exploitation Zone

Analysis-Ready Features (CSV)

Shannon Diversity Index

Multicultural diversity metric

H = -Σ(p_i × log(p_i))

- Nationality proportions
- Information theory basis
- By neighborhood/year
- Higher = more diverse
- Range: 0 to 3+
- CSV output format

Gini Coefficient

Income inequality measure

By district (inter-neighborhood)

- Rank by income index
- Window functions used
- Range: 0 (equal) to 1
- Additional statistics:
  - Mean, median, std dev
  - Coefficient of variation

Resolution Time KPI

Service efficiency metric

Days from creation to closure

- Date arithmetic logic
- Average by neighborhood
- Min/max values tracked
- Incident count included
- Quality issue found: 16.4% same-day resolution

Technical Implementation

Apache Spark 3.x | PySpark DataFrame API | Parquet Columnar Storage

Partition Pruning | Broadcast Join Optimization | Window Functions