Using Tags in Big Data Processing in the Tourism Industry and Wider Services Sphere to Overcome the Limitations of Traditional Categorisation

Olena KOT **olena.kot@hneu.net**

Nadiya DEKHTYAR **nadiya.dekhtyar @hneu.net**

Nadiia PROSKURNINA **nadiia.proskurnina@hneu.net**

Simon Kuznets Kharkiv National University of Economics

In line with the recommendations of UN Tourism (formerly UNWTO), the recording, analysis and evaluation of tourism activities are carried out in accordance with international standards, covering five logical modules

- **1. Classification of forms and types of tourism** - the basic categories of tourism according to geographical criteria, as well as specific types of tourist activity depending on the main purpose of the trip, the visitor's motivation, and how leisure time is spent.
- **2. Travellers, visitors and their environment** - a conceptual framework for classifying participants in tourism activities and their social groups, as well as the spatial and temporal criteria for determining visitor status.
- **3. Tourism products (goods and services), the tourism sector, and destination management (recreational areas)** - the structure of the tourism offering, including the development of tourism products, how destinations function and the organisations that manage them, the value chain and the labour market.
- **4. The economics of tourism and the composition of travel expenditure**: indicators for the macroeconomic measurement of tourism; the statistical recording of tourism expenditure and consumption; and a standardised, international tool - the tourism satellite account (TSA).
- **5. The methodology of statistical surveys and methods of ensuring data quality,** including statistical tools, data types, units of observation, survey methods, and criteria for assessing the quality of statistical information.

- substantive blocks:
  1. Classification of forms and types of tourism
  2.Travellers, visitors and their environment
- functional blocks:
  4. The economics of tourism and the composition of travel expenditure
  5. The methodology of statistical surveys and methods of ensuring data quality
- block three **3. Tourism products (goods and services), the tourism sector, and destination management (recreational areas)** serves as a bridge
- A lack of clarity in the substantive sections undermines the transparency and comprehensibility of the statistical observation system. This makes it impossible to reliably assess the impact of tourism on national and regional economies.

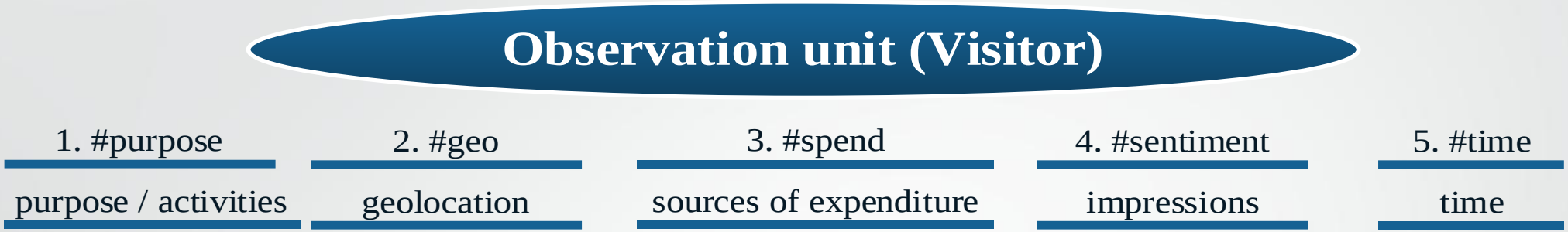
Key principles and methods for metadata tagging

Aspect / Category	Origins and evolution	Use in big data analysis	Free / open-source software
Tagging resources in clouds	In the late 2000s (introduced around 2008–2010 by platforms such as AWS)	Organising cloud servers, storage and databases by project, department or cost centre.	OpenTofu / Terraform, AWS CLI / LocalStack
Tags in data lakes and data catalogues	In the mid-to-late 2010s (developed alongside frameworks such as Apache Hive and catalogues such as Apache Atlas).	Managing petabytes of structured and unstructured data beyond the scope of simple file structures.	Apache Atlas (data catalogue), Apache Hive
Automated data search	Modern data catalogues and search interfaces.	Analysts use tags (e.g. #type) to quickly filter and search for the required data within massive data lakes without manual verification.	Amundsen від Lyft), DataHub (Open-Source catalogue by LinkedIn)
Data management and compliance	Provided through automated security policies and data platforms.	Tags identify sensitive information (e.g. #type:pii, classification=confidential) for automatic data masking or blocking unauthorised queries.	Apache Ranger (a security tool with support for tag-based policies)
Targeted processing and ETL (extract, transform, load)	Integrated into data loading and transformation pipelines.	Big data processing tasks (such as Apache Spark) read object tags to dynamically route and process data based on their labels (e.g., #type:raw_logs).	Apache Spark (data processing framework), Apache Airflow / Dagster (orchestrators)
Data lineage and change impact analysis	Supported by data tracing tools and cataloguing software.	Tags indicate an object's source or business domain, enabling engineers to trace how changes at higher stages (upstream) affect models and dashboards further down the chain (downstream).	OpenLineage / Marquez (standard and data tracing server), Spline (tracing for Spark)
Cost allocation and control	Used in cloud-based data storage and processing environments.	Tags (e.g. #cost-centre:marketing) track and audit specific storage and computing costs by individual business units.	Infracost (Estimating cloud costs by tag), Kubecost

Modern ecosystem of open-source and free tools for:
- data management
- automated data pipelines
- cost control in cloud-based data lakes

- Apache Atlas- deep integration with enterprise data stores, enabling the detailed classification of objects and the creation of rich metadata systems
- LinkedIn's DataHub & Lyft's Amundsen - search interfaces, automatic indexing of data sets, filtering resources instantly by tags
- Apache Ranger - centralised management of access policies based on tags (tag-based authorisation), that dynamically applies masking rules or completely block unauthorised queries to analytical databases
- Apache Spark + Apache Airflow + Dagster orchestrators - direct data processing and transformation pipeline construction (ETL)
- OpenTofu, LocalStack, Infracost, Kubecost - deployment and cost analysis tools

- Tagging systems operate in combination with five main categories



Group 1. Types of holiday and purpose of travel (#purpose/#activity):

#purpose:business – business tourism, professional trips or fact-finding visits aimed at expanding one's own business into the market of the country being visited, **#purpose:leisure** – holidays, leisure and relaxation, **#purpose:education** – educational tourism or a short-term academic exchange, **#activity:museum_visit** – visiting local museums and cultural attractions, **#activity:conference** – participation in professional conferences, exhibitions or corporate events.

Group 2. Geolocation and spatial routes (#geo/#route): **#geo:city_A** – reference to a specific settlement or destination;

#geo:recreational_zone: a recreational area or geographical zone with a high concentration of attractions;

#route:multi_destination: a tourist following several diverse routes.

Group 3. Sources of travel funding and expenditure channels (#spend):

#spend:hotel_booking: expenditure on accommodation and booking hotel services; **#spend:transit_tickets**: purchase of tickets for various modes of transport; **#spend:excursions** – booking and purchasing excursion programmes;

#spend:insurance – purchasing an insurance policy.

Group 4. Impressions and emotional analysis (#sentiment/#feedback):

#sentiment:positive/negative – the emotional tone identified through the analysis of social media posts or surveys;

#feedback:hotel_service_issue: a tag for recording complaints about the quality of hotel services.

Group 5. Time characteristics (#time/#duration): **#time:1_week** – length of stay at the destination (up to one week, for example); **#time:timestamp** – timestamp of a specific visit to track peak visitor numbers.

base data collection and 'tourist trail' tracking mechanisms on automatic tagging

- 1) **the City Card** (a visitor card offering bonuses and discounts) to track tourists' movements along routes and at stops when visiting attractions and using public and private transport;
- **2) reports from hotels and other accommodation providers**, as well as **leisure organisers**, via automated data aggregation from service providers and organisers of business and entertainment events;
- **3) surveys and social media monitoring**, achieved by including short questionnaires and analysing visitor feedback text

Question from the practice

- *in natural language:* 'Calculate the total expenditure of tourists who stayed in City A for one week, arrived on business, visited at least two museums, and left a negative re-view about the quality of service at their hotel'.

The filtering algorithm in the data lake would look like this:

$$\text{Filter} = (\#\text{geo: city_A}) \wedge (\#\text{time: 1_week}) \wedge (\#\text{purpose: business}) \wedge (\#\text{activity: museum_visit} \geq 2) \wedge (\#\text{sentiment: negative} \wedge \#\text{feedback: hotel_service_issue})$$

To ensure the development of a user-friendly and cost-effective environment for analysing shared data lakes, **a centralised regulator** (for example, the Ministry of Digital Transformation of Ukraine, in collaboration with experts from the National Tourism Organisation of Ukraine) should play a key role in supporting this system.

a hybrid (polyglot) non-relational architecture

- a document-oriented database
- indexing/caching tools

in order to collect and process analytical data on tourist trips based on multidimensional tagging sources and eliminate the problem of double counting

- Each observation unit (a tourist's trip or visit) is stored as a flexible, semi-structured document (JSON/BSON) in the system
- new tag categories are added dynamically without altering the entire database schema
- Each document represents a single trip or visitor ('tag carrier') and contains a comprehensive set of multidimensional tags


```
{
  "_id": "trip_2026_987654",
  "visitor_id": "usr_anon_88321",
  "data_source": "CityCard_Integration",
  "time_context": {
    "start_date": "2026-09-10T08:00:00Z",
    "end_date": "2026-09-17T18:00:00Z",
    "duration_days": 7
  },
  "geo_context": {
    "primary_destination": "Місто А",
    "recreational_zones": ["center_historic", "museum_district"]
  },
  "tags": {
    "purpose": ["business", "professional_event"],
    "activities": [
      { "type": "museum_visit", "count": 2, "locations": ["national_art",
"city_history"] },
      { "type": "conference_attendance", "count": 1 }
    ],
    "spend_breakdown": {
      "total_amount": 18500.00,
      "currency": "UAH",
      "channels": ["hotel_booking", "transit_tickets", "excursions"]
    },
    "sentiment": {
      "overall": "negative",
      "feedback_tags": ["hotel_service_issue"]
    }
  },
  "governance": {
    "classification": "pii_masked",
    "retention_policy": "standard_3_years"
  }
}
```

- A diagram of data collection and analysis based on three key NoSQL layers

DATA SOURCES
(City Card, information from accommodation providers, questionnaires and surveys)

↓

1. INPUT AND CACHING LAYER (In-Memory Key-Value: Redis / Aerospike)
- Rapid recording of the ‘tourist trail’ (City Card)
- Real-time deduplication of connections

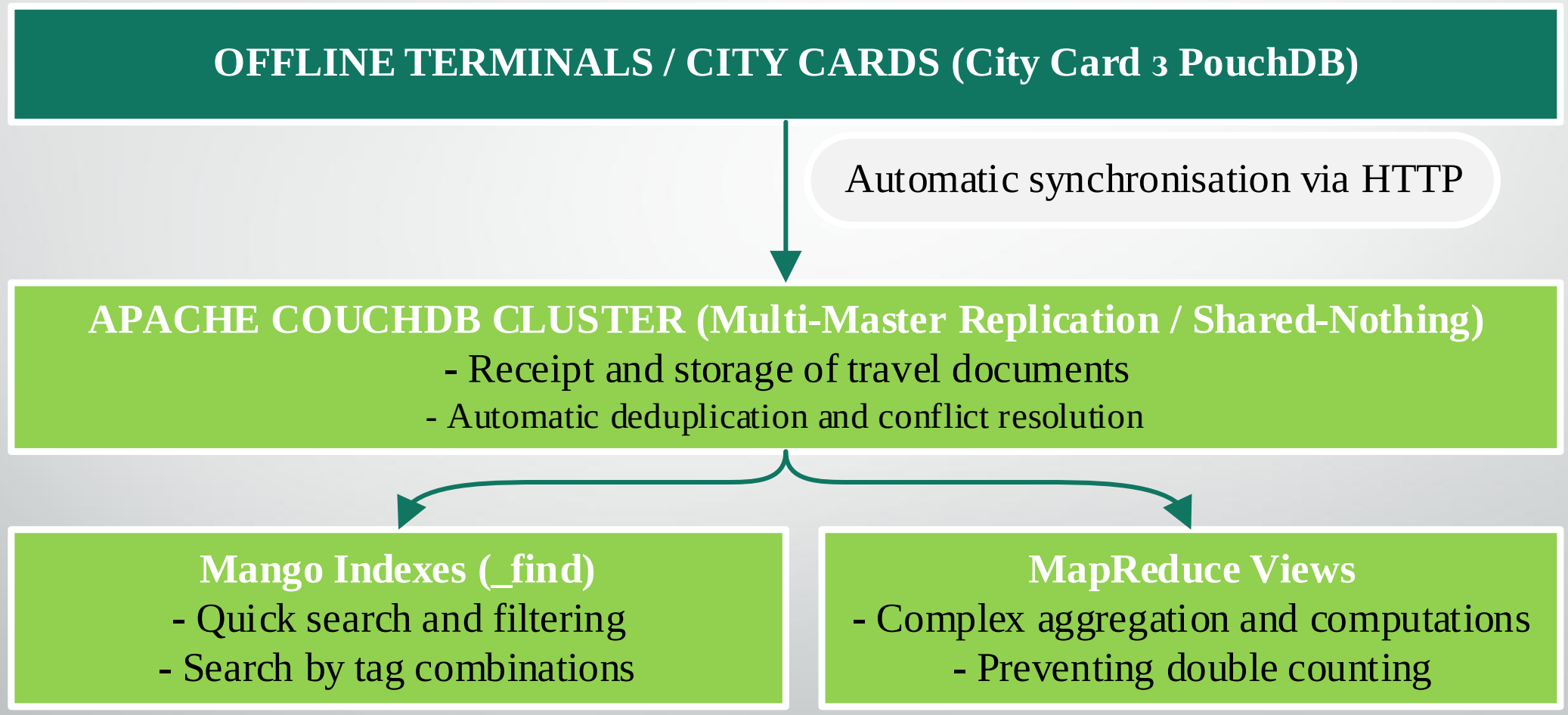
↓

2. MAIN DOCUMENT ARCHIVE (Document Store: MongoDB / Couchbase)
- Preserving the complete observation unit with tag arrays
- Multi-key indexing

↓

3. SEARCH AND ANALYTICS LAYER (Search Engine / Data Lake Integration)
- Elasticsearch / OpenSearch (for automated data search by tags)
- Експорт у Parquet / Delta Lake for processing via Apache Spark

- The overall architecture for data collection in Apache CouchDB



- *A shortcut:* the Map function analyses each document and selects only those that meet all the conditions, returning a unique visitor_id and the total cost.

```
function (doc) {
  if (doc.type === 'tourism_trip' && doc.geo_context && doc.tags) {
    // Checking geographical and temporal conditions
    var isTargetLocation = doc.geo_context.primary_destination === 'Місто A' && doc.time_context.duration_days === 7;
    // Tags' check
    var isBusiness = doc.tags.purpose && doc.tags.purpose.indexOf('business') !== -1;
    var isNegative = doc.tags.sentiment && doc.tags.sentiment.overall === 'negative';
    var isHotelIssue = doc.tags.sentiment && doc.tags.sentiment.feedback_tags && doc.tags.sentiment.feedback_tags.indexOf('hotel_service_issue') !== -1;
    // Checking activities (museums >= 2)
    var hasMuseums = false;
    if (doc.tags.activities) {
      for (var i = 0; i < doc.tags.activities.length; i++) {
        if (doc.tags.activities[i].type === 'museum_visit' && doc.tags.activities[i].count >= 2) {
          hasMuseums = true;
          break;
        }
      }
    }
    if (isTargetLocation && isBusiness && isNegative && isHotelIssue && hasMuseums) {
      // The key is 'visitor_id', and the value is 'expenditure'
      emit(doc.visitor_id, doc.tags.spend_breakdown.total_amount);
    }
  }
}
```

- The maximum volume of data that can be stored and processed in such a system depends on its two-tier architecture, which separates the in-memory NoSQL store (CouchDB) used for ingesting data from the analytical data lake (Spark/Delta Lake) used for in-depth analytics.

Number of registered journeys	Total number of tags attached (~10 tags per journey)	Storage size in CouchDB, including indexes	Volume in a data lake (Parquet / Delta Lake)	Optimal tool for processing
1 million	10 million	~2 – 3 Gb	~250 – 300 Mb	Apache CouchDB / PostgreSQL (jsonb)
10 million	100 million	~20 – 30 Gb	~2.5 – 3 Gb	CouchDB / Single-node Spark
100 million	1 milliard	~200 – 300 Gb	~25 – 30 Gb	Cluster CouchDB + Apache Spark
1 milliard	10 milliard	~2 – 3 Tb	~250 – 300 Gb	Distributed Data Lake (Spark / Trino)

Spark can perform full filtering and aggregation of **1 billion trips** (approximately **10 billion tags**) in just **10-30 seconds** in a standard cluster environment

- **Decoupling physical observation units** ('label bearers') **from assigned multidimensional operational tags** addresses a fundamental methodological bottleneck in traditional statistical classification models.
- By maintaining a single observation unit to which a matrix of dynamic tags is attached, the proposed framework **guarantees mathematical de-duplication** during regional headcount calculations, while preserving tag cardinality for assessing attraction popularity.
- Incorporating **a proportional expenditure allocation function** driven by spatio-temporal interest density allows visitor budgets to be distributed across visited nodes without inflating regional totals artificially.
- This mathematical reconciliation ensures that micro-level telemetry feeds are **fully aligned with national accounting invariants and macro-level frameworks**, such as the UN Tourism Satellite Account