Journalism Cloud Alliance Inaugural Meeting

17 April 2024

SURVEY FINDINGS
ON CLOUD AND DATA PROCESSING SERVICES FOR NEWSROOMS

SURVEY:
Methodology and sample

The survey was designed to explore how investigative journalism organisations and data journalism newsrooms harness data processing tools and cloud services. **Structured questionnaires** were sent to a diverse array of **journalism organisations**, with **18 responding during January and February 2024**.

The survey also examined financial commitments to these technologies, satisfaction with cloud services, strategies for AI and machine learning, and interest in forming coalitions for better technology deals. The respondents ranged from **small teams of fewer than 10 journalists to large groups of over 100**, reflecting a spectrum of practices and challenges in data processing.

SURVEY:
Methodology limitations

The survey's findings are based on a sample of **18 investigative journalism organisations** and **data journalism newsrooms**.

- The structured questionnaire approach relies heavily on the **accuracy and completeness of the responses provided by the participants**.

- The data was collected during **January and February 2024**, which might not capture long-term trends or changes in organisational practices and technology usage that could develop over a longer period.

- The survey includes **financial expenditures** as reported by the organisations, which could vary significantly in accuracy depending on the internal financial tracking and reporting standards of each organisation.

Organisational structures and expenditures

- **Organisational Size:**

  - **39%** of newsrooms have **fewer than 10 journalists**.

  - 28% have 10-50 full-time journalists in their organisations.

  - 22% between 51 and 100 and 11% over 100 journalists.

  - **12%** have **over 100 journalists**.

Organisational structures and expenditures

- **Staffing Costs:**
  - **53%** allocate **less than $5,000 monthly** to staff managing data technology.

  - **30%** of respondents indicated their salary expenditures fall **within the $5,000 - $10,000 range**

  - **11%** indicated a monthly expenditure ranging **between $10,000 and $20,000**.

  - The estimated **total annual spending on salaries** among all respondents **is approximately $1,080,000 USD**.

Cloud Services

- **Amazon Web Services (AWS):**
  - Frequently mentioned with services like EC2 and S3.
  - Noted for a range of applications from general storage to specific uses in ML/AI.
- **Google Cloud Platform (GCP):**
  - Cited for both storage and compute services.
  - Examples include "GCP Storage" and "GCP Compute".
- **Digital Ocean:**
  - Mentioned for hosting services like Droplets.
  - Noted for ease of use and affordability.
- **FlokiNet and GreenHost:**
  - These providers are mentioned for specific types of hosting such as "Bare metal" and VPS hosting.
  - Indicated use by organizations with specific privacy or ethical hosting requirements.

Toolkits

- **Self-built and open-source tools:**
  - Datashare and Aleph that provide data functions like data browsing, access to large data archives and collaboration.
- **Business, analytics and support software:**
  - Clarity app (user behaviour analytics), Tunnel Bear (secure browsing), Torrent Tracker (p2p file sharing) and Discourse Safe Box (secure communications).
- **Infrastructure tools:**
  - OAuth 2 (secure authorisation protocols), Grafana and Prometheus (real-time server usage statistics) and Watchtower (monitoring).
  - Each tool serves a unique purpose and forms part of a stack allowing sharing and processing of massive datasets.

Expenditures on Data Storage & Processing

Current Spending on Data Systems
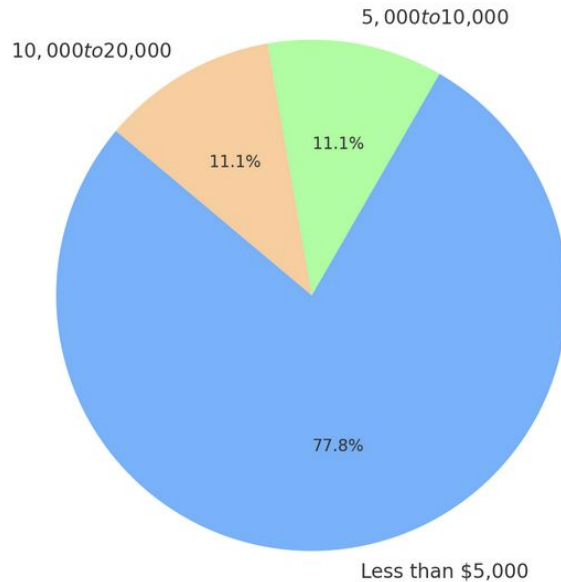
Future Spending Expectations on Data Systems
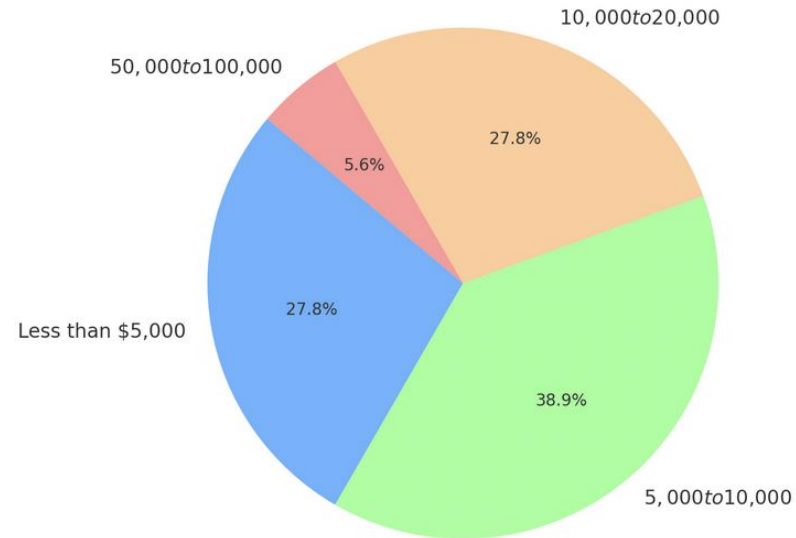
Expenditures on Data Storage & Processing

- Current Monthly Spending:
  - **78%** spend **less than $5,000.**
  - **11%** spend between **$5,000 and $10,000.**
  - **11%** spend between **$10,000 and $20,000.**

- Future Spending Projections:
  - **28%** expect to spend less than **$5,000**.
  - **39%** expect spend between **$5,000 to $10,000**.
  - **28%** expect spend between **$10,000 to $20,000**.
  - **5%** expect spend between **$50,000 to $100,000**.

Organisational structures and expenditures

- **Role of Large Volumes of Data, Audio, Video, and Imagery:**
  - All organisations **expect large volumes of data storing and processing to play an increasingly larger role** in their activities over the next 2 to 3 years.
  - Similarly, respondents anticipate a **greater role for audio, video, and large imagery** (including satellite imagery) in their future operations.

Expenditures on Data Storage & Processing

The estimated total annual spending on data services, storage, and processing systems, as **expected 2 to 3 years from now** by the surveyed organizations, is **approximately $2,580,000 USD**.

The estimated total **current annual spending** on data storage and processing systems for all respondents is **approximately $960,000 USD**.

Preparing for tomorrow: capacity building and technological adaptation

*"In the forthcoming two to three years, we anticipate a **complex and evolving landscape** in data storage and processing. As the volume of data increases**, security and privacy issues become more pressing, necessitating stronger protection measures and adherence to regulatory standards**.*
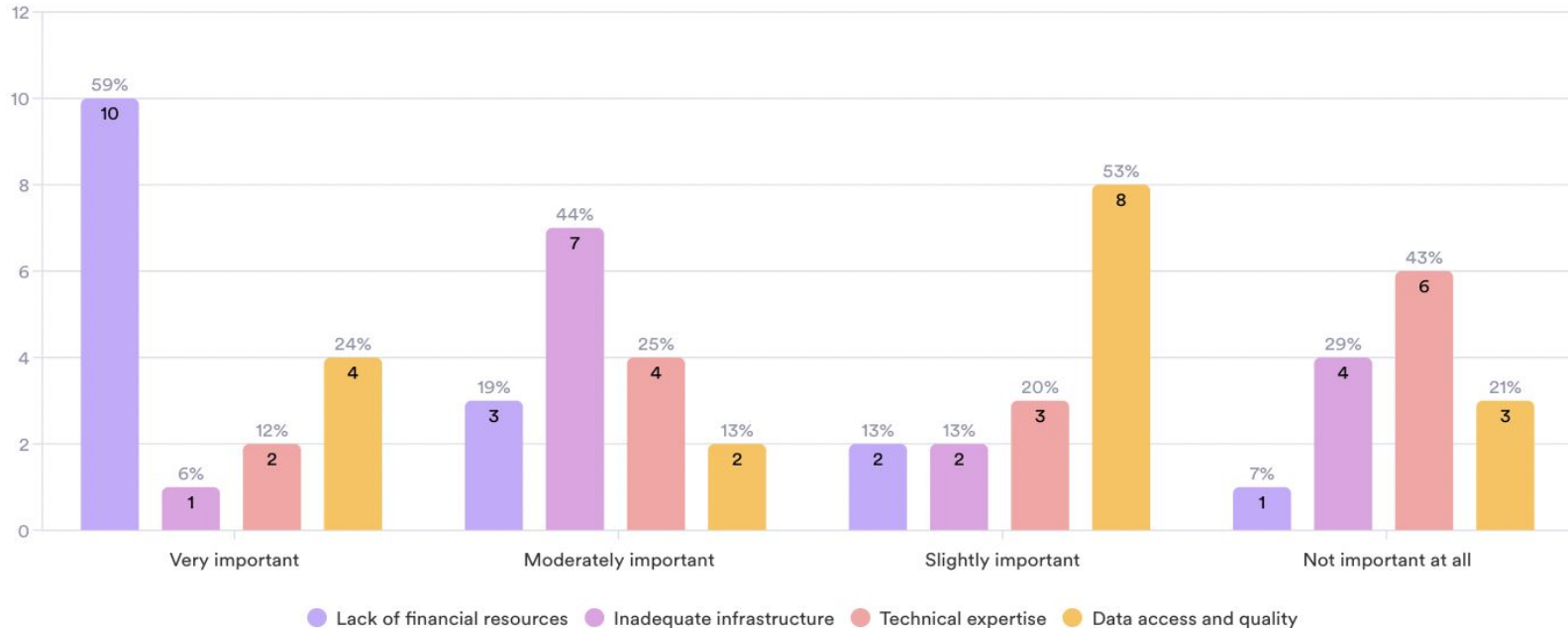
*Scalability and performance optimisation remain critical for managing growing datasets. The **integration of artificial intelligence and machine learning poses new requirements for storage and processing infrastructures**."*

Survey respondent

Challenges of data management

5. What are the primary challenges or obstacles your organisation faces in carrying out journalism work that requires increasingly large volumes of data? (Rank the challenges in order of impact)

17 Responses

Legend: Lack of financial resources · Inadequate infrastructure · Technical expertise · Data access and quality
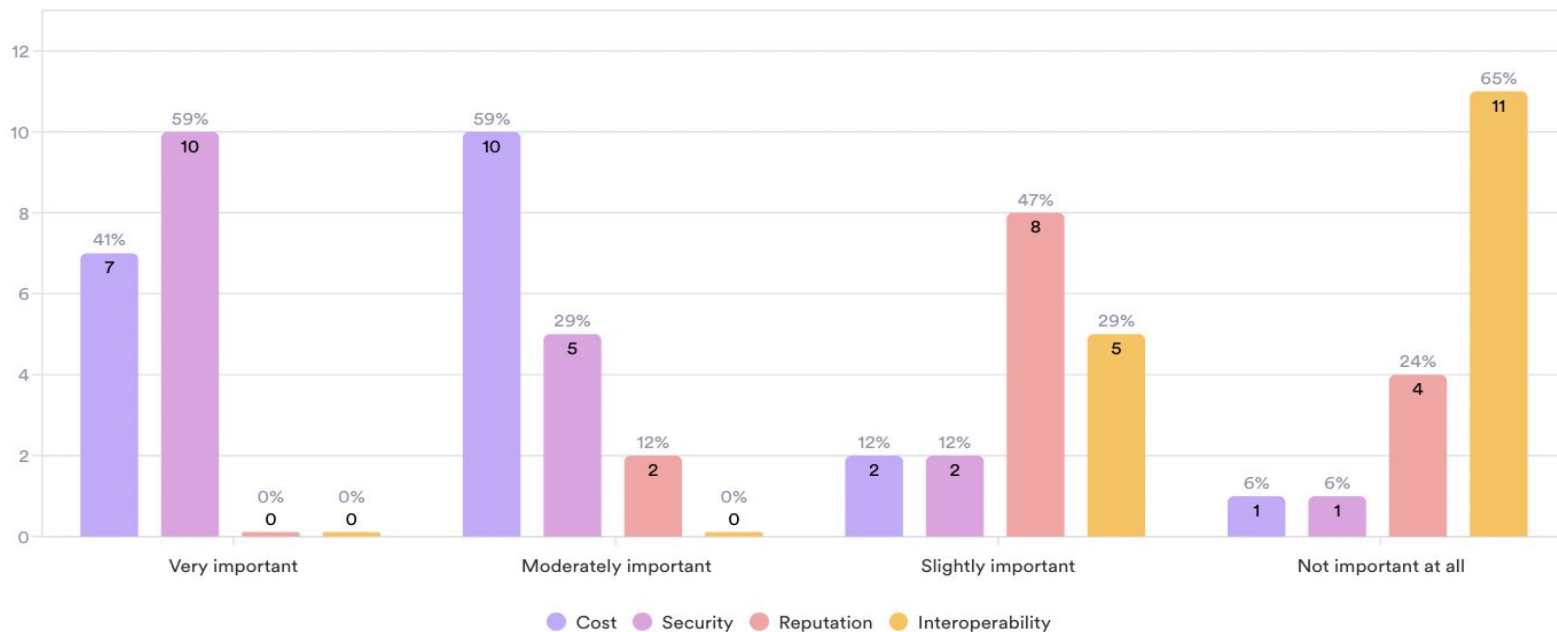
Computational and cloud infrastructure

- **Inadequate Infrastructure:**
  - **53%** of respondents feel their **current computational and cloud infrastructure is insufficient** for future journalism work and investigations planned for the next 2 to 3 years.

Computational and cloud infrastructure: current state and priorities

GFMD
GLOBAL
FORUM
FOR MEDIA
DEVELOPMENT

10. How important are the following considerations in deciding which computational and cloud services to use?

17 Responses

Computational and cloud infrastructure

- **Security concerns:**
  - All but one respondent confirmed that **they share concerns about the security of confidential and sensitive data** when it is stored with cloud providers that are also developing their in-house AI foundation models.

Computational and cloud infrastructure: current state and priorities

- **Top Concerns in Service Selection:**
  - **Safety** is a predominant issue with **59%** rating it as **"very important"** and **29%** as **"moderately important"**. Concerns are particularly pronounced regarding the security of sensitive data with cloud providers that develop their own AI models.
  - **"Cost Considerations"** as a significant factor with **41%** considering it 'very important' and **59%** as 'moderately important'.
  - **Reputation is a minor factor** (**47%** slightly important), and **interoperability** is generally not a significant concern (**65%** not significant).
- **Provider Switching Barriers:**
  - A notable **76%** of respondents have never switched cloud providers, suggesting significant barriers due to the complex nature and technical variations of cloud services.

AI and machine learning: a new frontier in journalism

- Nearly **half of the respondents** plan to **engage with AI and ML technologies** within the next **2 to 3 years**, indicating the growing relevance of these technologies in journalism.
- **Applications and Intentions**:
    - Fact-Checking
    - Document Processing
    - Support for Media Professionals
- **Development and Security Concerns**:
    - Some journalists are exploring algorithms used by public bodies.
    - Others are developing tools using existing APIs specifically tailored to journalistic needs.
    - Local Language Tools: A few respondents have created tools in local languages and are focusing on tools that ensure robust security without compromising functionality.

Modeling Future Costs

- Assuming a plan to fully translate a dataset of **six terabytes** with translation models at **current pricing** would be quickly cost prohibitive.
- **A real-life:**
  - Most corporate email inboxes can hold **10Gb of data** which could hold ~480.000 inboxes in **six terabytes**
  - Amazon current charges ~0,05 USD *per* email translation
  - In one estimate, the respondents calculated **over ~$200.000USD** to translate **one dataset**.

Collaboration for Data-intensive Future

There is a strong **consensus** among respondents on the **need for collaboration among journalism organizations to navigate a data-intensive future effectively**.

- **Benefits of Collaboration**:
    - Better Deals and Support
    - Capacity Building
    - Technical and Financial Assistance
    - Secure Tool Adoption: Interest in identifying secure tools and best practices, with a proactive approach to engaging with providers to protect user rights.
- **Engagement and Learning**:
    - Willingness to engage in detailed discussions with a technology alliance, highlighting a proactive attitude towards learning and adapting to technological advances.

Conclusions & Recommendations

Journalism organisations show a **wide range of monthly spending** on **data management**, from under **£5,000** to **£20,000**.

1

Conclusions & Recommendations

Many newsrooms plan to **double their investment** in advanced data processing technologies within the next **two** to **three years**.

2

Conclusions & Recommendations

Essential **digital tools for data management** are widely used across **newsrooms** for various functions, including **security** and **analytics**.

3

Conclusions & Recommendations

Security is the **primary concern** when selecting cloud services, driven by the need to protect sensitive data.

4

Nearly half of the surveyed organisations plan to adopt **AI** and **machine learning** soon, especially for document processing and fact-checking.

5

Conclusions & Recommendations

There is a **unanimous view** on the benefits of collaboration for securing **better technology deals** and **resource sharing**.

6

Conclusions & Recommendations

Organisations are preparing for **future challenges** with data **security**, **privacy**, and **cloud management**, necessitating strategic technology adoption.

7

Thank you!

For more information please contact:

Mira Milosevic

Executive Director

mmilosevic@gfmd.info

Find additional information about the meeting and related resources at the GFMD IMPACT resource centre